



Munich Personal RePEc Archive

Heterogeneous component multiplicative error models for forecasting trading volumes

Antonio Naimoli and Giuseppe Storti

University of Salerno, University of Salerno

9 May 2019

Online at <https://mpra.ub.uni-muenchen.de/93802/>

MPRA Paper No. 93802, posted 10 May 2019 01:57 UTC

Heterogeneous component multiplicative error models for forecasting trading volumes

Antonio Naimoli, Giuseppe Storti*

Università di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES)

Abstract

We propose a novel approach to modelling and forecasting high frequency trading volumes. The new model extends the Component Multiplicative Error Model of [Brownlees et al. \(2011\)](#) by introducing a more flexible specification of the long-run component. This uses an additive cascade of MIDAS polynomial filters, moving at different frequencies, in order to reproduce the changing long-run level and the persistent autocorrelation structure of high frequency trading volumes. After investigating its statistical properties, the merits of the proposed approach are illustrated by means of an application to six stocks traded on the XETRA market in the German Stock Exchange.

Keywords: Intra-daily trading volume, dynamic component models, long-range dependence, forecasting.

1. Introduction

Thanks to the rapid growth of computing power and availability of data storage facilities, in recent years, the analysis of high frequency data has been receiving increasing attention in the financial econometrics literature. The availability of financial data recorded at very high frequencies has inspired the development of new types of econometric models, able to reproduce the peculiar features of these data such as strong serial dependencies, irregular spacing in time, price discreteness and intra-daily seasonal patterns. To model the dynamic behaviour of irregularly spaced transaction data, [Engle and Russell \(1998\)](#) proposed the Autoregressive Conditional Duration (ACD) model, later generalised in Multiplicative Error Model (MEM) by [Engle \(2002\)](#). MEMs are a general class of time series models for positive-valued random variables, which are decomposed into the product of their conditional mean and a positive-valued error term with unit mean. Extensions of this class of models and their statistical properties are discussed in [Chou \(2005\)](#), [Manganelli \(2005\)](#), [Cipollini et al. \(2006, 2013\)](#), [Lanne \(2006\)](#), [Brunetti and Lildholdt \(2007\)](#) and [Brownlees et al. \(2011\)](#), among others.

At the same time, the continuous development of new financial instruments has stimulated the research on statistical models for positive-valued time series, such as number of trades and volumes, high-low range, absolute returns, financial durations and realized volatility measures derived from ultra high frequency data. It is well established that all these variables have rich serial dependence structures sharing the features of clustering and high persistence. The recurrent feature of long-range dependence is conventionally modelled through the use of autoregressive fractionally integrated moving average (ARFIMA) models, as in [Andersen et al. \(2003\)](#), or using regression models mixing information at different frequencies, such as the Heterogeneous AR (HAR) model of [Corsi \(2009\)](#). The HAR model, initially proposed for modelling realized volatility series, is inspired by the Heterogeneous Market Hypothesis of [Müller et al. \(1993\)](#) and offers a simple alternative to the use of ARFIMA models. The idea is to model daily realized volatility as a linear combination of past realized volatilities aggregated at different frequencies. Despite its simplicity, in practical applications, this particular structure, usually referred to as an *additive volatility cascade*, has been

*Corresponding author: Giuseppe Storti, Università di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES), Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy. *Email address:* storti@unisa.it.

found to be able to satisfactorily reproduce the empirical regularities of realized volatility series, including their highly persistent autocorrelation structure.

It is worth noting that, working with real data, spurious persistence could also occur as a consequence of unmodelled structural breaks or level shifts. These phenomena could take place smoothly or in abrupt manner. In order to model changes in the long-run level of the variable of interest, still remaining within the class of MEMs, [Gallo and Otranto \(2015\)](#) proposed a new class of models that combine Markov switching models with smooth transition dynamics, with the aim of keeping track of both smooth and abrupt level changes. Along the same line, researchers and practitioners have recently shown interest for *component models* featuring two or more components moving at different frequencies. The increasing popularity of these models is due to their ability to parsimoniously characterise the rich dependence structure of financial variables such as volatility and volume. Component models have been initially applied to daily returns in a GARCH framework. Starting from the Spline GARCH of [Engle and Rangel \(2008\)](#), where volatility is specified as the product of a slow-moving component, represented by an exponential spline, and a short-run component, following a unit GARCH process, several contributions have extended and refined this idea. [Engle et al. \(2013\)](#) introduced a new class of models, called GARCH-MIDAS, where the long-run component is modelled as a MIDAS (Mixed-Data Sampling, [Ghysels et al. \(2007\)](#)) filter that applies to monthly, quarterly or biannual financial and macroeconomic variables. [Brownlees and Gallo \(2010\)](#) proposed a dynamic model incorporating a long-run component based on some linear basis expansion of time and bounded with a penalised maximum likelihood estimation strategy. [Amado and Teräsvirta \(2013\)](#) decomposed the variance into a conditional and an unconditional component, so that the latter smoothly evolves over time through a linear combination of logistic transition functions taking time as the transition variable. A recent review of component volatility models can be found in [Amado et al. \(2019\)](#).

Moving to the analysis of intra-daily data, [Engle and Sokalska \(2012\)](#) developed the Multiplicative Component GARCH. This model decomposes the volatility of high frequency asset returns into the product of three components: a daily, a diurnal and a stochastic intraday component. In a multivariate setting, [Bauwens et al. \(2016\)](#) and [Bauwens et al. \(2017\)](#) developed and discussed several component specifications for time series of realized covariance matrices.

Applications of component models to the analysis of other positive-valued time series are more rare. Coming closer to the object of this paper, [Brownlees et al. \(2011\)](#) proposed a Component MEM (CMEM) for intra-daily trading volumes, where long-run (daily) and non-periodic short-run (intra-daily) dynamics are modelled using GARCH-type recursions moving at different frequencies. In order to account and test for asymmetric dynamics, the model includes a dummy, whose value depends on the sign of past stock returns. The specification is completed by a periodic component accounting for intra-daily seasonality. This component structure is found to be able to capture the salient features of intra-daily volumes such as high-persistence, asymmetry and intra-daily periodicity.

Aim of this paper is to propose novel dynamic component models for high frequency trading volumes, investigate their statistical properties and assess their effectiveness for trading by means of an out-of-sample forecasting exercise. The main specification proposed in this paper, called the Heterogeneous MIDAS Component Multiplicative Error Model (H-MIDAS-CMEM), is closely related to the class of Component MEMs discussed in [Brownlees et al. \(2011\)](#). The most notable difference with respect to the latter is that the long-run component is now modelled as an additive cascade of MIDAS filters moving at different frequencies (from which the *heterogeneous* quality of the model comes). This specification is motivated by the empirical regularities arising from the analysis of high frequency time series of trading volumes. After removing the intra-daily seasonal cycle, these are typically characterised by two prominent and related features: a slowly moving long-run level and a highly persistent autocorrelation structure. In our model, we account for both these features by considering a heterogeneous MIDAS specification of the long-run component. Residual short term autocorrelation is then explained by an intra-daily non-periodic component that follows a mean reverting unit GARCH-type process. In addition, from an economic point of view, the cascade structure of the long-run component reproduces the natural heterogeneity of financial markets that are typically characterised by different categories of agents operating at different frequencies. This results in a variety of sources separately affecting the variation of the average volume at various speeds.

Model parameters are estimated by a two stage approach. First, we correct the raw volumes for intra-day seasonality by a Fourier Flexible Form, whose coefficients are estimated by OLS regression. Then, the remaining parameters are estimated by the method of maximum likelihood under the assumption that the innovations are distributed according to the (Dynamic) Zero-Augmented Generalized F distribution introduced by [Hautsch et al. \(2014\)](#). The reason

for this choice is twofold. First, it delivers a flexible probabilistic model for the conditional distribution of volumes. Second, it allows to control for the presence of zero volumes in our data.

In order to assess the relative merits of the proposed approach we have performed a forecasting exercise using high frequency trading volume data from January 2009 to December 2012 for six stocks traded on the Xetra Market in the German Stock Exchange. The ability of the proposed models to predict intra-daily volumes at different horizons is assessed by means of purely statistical loss functions such as the Mean Squared Error (MSE) and Mean Absolute Error (MAE). In addition, in order to offer an economic appraisal of the proposed specifications, we have considered the Slicing loss function (Brownlees et al., 2011). This measures the effectiveness of volume forecasts for the implementation of trading strategies based on the replication of the Volume Weighted Average Price (VWAP). The results are compared with those generated by a set of alternative models, including the CMEM and its extension obtained specifying the long-run component of volumes as a HAR model (HAR-CMEM).

Our findings suggest that the H-MIDAS-CMEM is able to satisfactorily reproduce the salient empirical features of high frequency volumes. We also find that the forecasting performance of the H-MIDAS-CMEM favorably compares with that of its main competitors. The Model Confidence Set (MCS) of Hansen et al. (2011) is used to assess the significance of differences in the predictive performances of the models under analysis.

The remainder of the paper is structured as follows. Section 2 describes the proposed H-MIDAS-CMEM model defining its components. The statistical properties of the model are investigated in Section 3, where we provide conditions for strict stationarity and ergodicity of the seasonally adjusted volumes. The estimation procedure is presented in Section 4, while Section 5 illustrates the results of the empirical application. Section 6 concludes.

2. Model specification

Let $\{x_{t,i}\}$ be a time series of intra-daily trading volumes. We denote days by the subscript $t \in \{1, \dots, T\}$, where each day is divided into I equally spaced intervals indexed by $i \in \{1, \dots, I\}$. The total number of observations is then given by $N = TI$. In the remainder, it will be convenient to adopt the following convention: given non-negative integers j, w, k , with $w \leq I$, we let $x_{t,i-j} = x_{t-k,i-w}$ for $j = kI + w$.

The H-MIDAS-CMEM represents $x_{t,i}$ as the product of different stochastic components according to the equation

$$x_{t,i} = \tau_{t,i} g_{t,i} \phi_i \varepsilon_{t,i}, \quad (1)$$

where ϕ_i is an intra-daily periodic component that reproduces the approximately U-shaped intra-daily seasonal pattern typically characterising trading activity; $\tau_{t,i}$ is a smoothly varying component, given by the sum of MIDAS filters moving at different frequencies, designed to track the dynamics of the long-run level of trading volumes; $g_{t,i}$ is an intra-daily dynamic non-periodic component, based on a mean reverting unit GARCH-type process, that reproduces autocorrelated movements around the current long-run level. Finally, $\varepsilon_{t,i}$ is an error term satisfying the following assumption.

Assumption A1 (iid errors): *The multiplicative innovation term $\varepsilon_{t,i}$ is assumed to be an i.i.d. non-negative process with unit mean and constant variance σ^2 , that is*

$$\varepsilon_{t,i} \stackrel{iid}{\sim} \mathcal{D}^+(1, \sigma^2). \quad (2)$$

In the remainder of this section, we will discuss in more detail the structure of the dynamic components in Eq. (1). In addition, we will investigate the stochastic properties of the short-run non-periodic component $g_{t,i}$ and of the seasonal adjusted volumes $(x_{t,i}/\phi_i)$.

2.1. Intra-daily periodic component

Intra-daily volumes usually exhibit a U-shaped daily seasonal pattern, i.e. trading activity is higher at the beginning and at the end of the day than around lunch time. In order to model these periodicities, as in Engle and Sokalska (2012), we specify the intra-day seasonal component ϕ_i via a Fourier Flexible Form (Gallant, 1981)

$$\phi_i = \sum_{q=0}^Q a_{0,q} t^q + \sum_{p=1}^P \left[a_{c,p} \cos(2\pi p t) + a_{s,p} \sin(2\pi p t) \right], \quad (3)$$

where $\iota = i/I \in (0, 1]$ is a normalised intraday time trend and the number of terms (Q, P) in (3) can be selected through the use of standard information criteria. The diurnally adjusted trading volumes can be then computed as

$$y_{t,i} = \frac{x_{t,i}}{\phi_i}. \quad (4)$$

2.2. Intra-daily dynamic non-periodic component

As in [Engle et al. \(2013\)](#), the intra-daily non-periodic component $g_{t,i}$ is assumed to follow a mean reverting GARCH-type process with $E(g_{t,i}) = 1$. Namely, the dynamics of $g_{t,i}$ are determined by the following recursion

$$g_{t,i} = \omega^* + \sum_{j=1}^r \alpha_j \frac{y_{t,i-j}}{\tau_{t,i-j}} + \alpha_0 I(y_{t,i-1} = 0) + \sum_{k=1}^s \beta_k g_{t,i-k}, \quad \tau_{t,i} > 0 \quad \forall t, i, \quad (5)$$

where $I(y_{t,i-1} = 0)$ denotes an indicator function that takes value 1 if the argument is true and 0 otherwise. It is worth noting that Eq. (5) does not automatically guarantee that the unit mean assumption on $g_{t,i}$ is satisfied. To this purpose, it is necessary to set appropriate constraints on ω^* by means of a targeting procedure that will be discussed in Section 3.1.

The coefficient α_0 of the dummy variable in (5) has the role of adjusting the dynamics of $g_{t,i}$ to the lack of trading activity at time $(t, i - 1)$. This is most easily seen in the simple case in which $r = 1$ and $s = 1$, where Eq. (5) can be reformulated giving rise to the following piecewise linear model

$$\begin{cases} \forall y_{t,i-1} > 0, & g_{t,i} = \omega^* + \alpha_1 \frac{y_{t,i-1}}{\tau_{t,i-1}} + \beta_1 g_{t,i-1}, \\ \forall y_{t,i-1} = 0, & g_{t,i} = \omega^* + \alpha_0 + \beta_1 g_{t,i-1}. \end{cases}$$

In general, representing the dynamics of $g_{t,i}$ as a regime switching model offers a convenient framework for deriving sufficient conditions for the positivity of $g_{t,i}$ that, for ease of reference, are summarised in the following assumption.

Assumption A2 (positivity of $g_{t,i}$). *The parameters in (5) satisfy*

$$\omega^* > 0, \quad \alpha_j \geq 0 \quad (j = 1, \dots, r), \quad (\omega^* + \alpha_0) > 0, \quad \beta_k \geq 0 \quad (k = 1, \dots, s).$$

2.3. Low frequency component

The low frequency component is modelled as a linear combination of MIDAS filters of past volumes aggregated at different frequencies. It should be emphasised that our setting is more general than that considered by [Brownlees et al. \(2011\)](#) since we do not constrain the long-run component to be fixed within the trading day, but its value is updated as soon as a new intra-daily observation is made available. In addition, this choice facilitates the derivation of the stochastic properties of the seasonally adjusted volumes $y_{t,i}$ in Section 3.

In the MIDAS framework, a relevant issue is related to the identification of the frequency of the information to be used by the filter, that notoriously acts a smoothing parameter. A simple and intuitive solution would be to use volumes aggregated over a rolling window of daily length, leading to the following specification for $\tau_{t,i}$

$$\tau_{t,i} = m_d + \theta_d \sum_{k=1}^{K_d} \varphi_k(\omega_d) YD_{t,i}^{(k)}, \quad (6)$$

where

$$YD_{t,i}^{(k)} = \sum_{j=1}^I y_{t,i-(k-1)I-j}$$

denotes the rolling daily cumulative volume. The subscript d indicates that the parameters refer to the daily frequency. Furthermore, in order to guarantee the positivity of the estimated trend, we impose the following parameter constraints: $m_d > 0$ and $\theta_d > 0$. This does not imply any relevant loss of generality since traded volumes are typically positively autocorrelated.

A common choice for determining $\varphi_k(\omega_d)$ is the Beta weighting scheme

$$\varphi_k(\omega_d) = \frac{(k/K_d)^{\omega_{1,d}-1}(1-k/K_d)^{\omega_{2,d}-1}}{\sum_{j=1}^{K_d} (j/K_d)^{\omega_{1,d}-1}(1-j/K_d)^{\omega_{2,d}-1}}, \quad (7)$$

where the weights in Eq. (7) sum up to 1 and $\omega_d = (\omega_{1,d}, \omega_{2,d})'$. As discussed in Ghysels et al. (2007), this weighting function is very flexible, being able to accommodate increasing, decreasing or hump-shaped weighting schemes. Furthermore, the value of K_d , the number of daily lags involved in the filter, can be chosen by information criteria, to avoid overfitting problems.

An alternative trend specification could be based on the use of higher frequency volumes aggregated over intervals of length equal to $1/H$ days

$$\tau_{t,i} = m_h + \theta_h \sum_{k=1}^{K_d} \sum_{l=1}^H \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) YH_{t,i}^{(l,k)}, \quad (8)$$

where $n_h = I/H \in \{1, \dots, HT\}$ denotes what, for ease of reference, we will call the *hourly* period, H is the number of sub-intervals in which the day is divided, while the subscript h refers to the parameters corresponding to the just defined *hourly* frequency. The variable $YH_{t,i}^{(l,k)}$ corresponds to the (l) -th hourly cumulative volume of the (k) -th past interval of daily length before time i of day t . Namely,

$$YH_{t,i}^{(l,k)} = \sum_{j=1}^{n_h} y_{t,i-(k-1)I-(l-1)n_h-j}. \quad (9)$$

A more general formulation of the long-run component, encompassing the previous two, is then given by

$$\begin{aligned} \tau_{t,i} = m + \theta_d \sum_{k=1}^{K_d} \varphi_k(\omega_{1,d}, \omega_{2,d}) YD_{t,i}^{(k)} \\ + \theta_h \sum_{k=1}^{K_d} \sum_{l=1}^H \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) YH_{t,i}^{(l,k)}. \end{aligned} \quad (10)$$

This multiple frequency specification appears to be preferable to the previous single-frequency models in (6) and (8) for three different reasons. First, the modeller is not bound to choose a specific frequency for trend estimation, but can determine the optimal blend of low and high frequency information in a data driven fashion. Second, it is compatible with the heterogeneous market assumption of Müller et al. (1993), enforcing the idea that market agents can be divided in different groups characterised by different frequencies of interest and strategies. Third, as pointed out in Corsi (2009), an additive cascade of linear filters, applied to the same variable aggregated over different time intervals, can allow to reproduce very persistent dynamics such as those typically observed for high frequency trading volumes.

Positivity of the long-run component $\tau_{t,i}$ can be guaranteed by imposing appropriate constraints on the parameters of (10). These are summarised in the following assumption.

Assumption A3 (positivity of $\tau_{t,i}$). Assume that the parameters in (10) satisfy

$$m > 0, \quad \theta_d > 0, \quad \theta_h > 0, \quad \varphi_k(\omega_{1,d}, \omega_{2,d}) \geq 0, \quad \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) \geq 0,$$

$\forall l, k$, with

$$\sum_{k=1}^{K_d} \varphi_k(\omega_{1,d}, \omega_{2,d}) = 1, \quad \sum_{k=1}^{K_d} \sum_{l=1}^H \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) = 1.$$

3. Dynamic properties of seasonally adjusted volumes

3.1. Statistical properties of the short term component $g_{t,i}$

In this section, we provide conditions under which the short-run component $g_{t,i}$ is strictly stationary and ergodic. Before proceeding with the illustration of the main results, we define some notational conventions that will be used throughout this section. In particular, we consider the following norm: for $\mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\| = \max_{i=1,2,\dots,n} (v_1, \dots, v_n)'$ and, for a $(m \times n)$ matrix M , the induced matrix norm is $\|M\| = \max_{i=1,\dots,n} \sum_{j=1}^n |M_{ij}|$. Also, \mathbf{I}_n denotes an identity matrix of order n ; $\mathbf{0}_{m,n}$ denotes a $(m \times n)$ matrix of zeros; $\mathbf{1}_{m,n}$ denotes a matrix of ones of the same dimension; $\log^+(x) = \max(\log(x), 0)$.

The desired result is obtained following the approach described in [Bougerol and Picard \(1992a\)](#). First, we need to rewrite Eq. (5) as a random coefficient vector autoregressive model

$$\mathbf{g}_{t,i} = B_{t,i} + A_{t,i}\mathbf{g}_{t,i-1}, \quad (11)$$

where $\mathbf{g}_{t,i}$ is the $(r + s - 1) \times 1$ stochastic vector

$$\mathbf{g}_{t,i} = [g_{t,i}, \dots, g_{t,i-s+1}, \tilde{y}_{t,i-1}, \dots, \tilde{y}_{t,i-r+1}]'.$$

with $\tilde{y}_{t,i} = y_{t,i}/\tau_{t,i}$ ($\tau_{t,i} > 0$). Without loss of generality, we impose $r, s \geq 2$, that can always be obtained even from lower order models by introducing some additional α and β coefficients equal to 0, if needed. $A_{t,i}$ is a positive-valued random matrix that can be written in block form as

$$A_{t,i} = \begin{pmatrix} \delta_{t,i} & \beta_s & \alpha & \alpha_r \\ \mathbf{I}_{s-1} & 0 & \mathbf{0}_{s-1,r-2} & 0 \\ \xi_{t,i} & 0 & \mathbf{0}_{1,r-2} & 0 \\ \mathbf{0}_{r-2,s-1} & 0 & \mathbf{I}_{r-2} & 0 \end{pmatrix},$$

where $\delta_{t,i}$ is a $1 \times (s - 1)$ random vector

$$\delta_{t,i} = [\beta_1 + \alpha_1 \varepsilon_{t,i-1}, \beta_2, \dots, \beta_{s-1}],$$

$\xi_{t,i}$ is a $1 \times (s - 1)$ random vector

$$\xi_{t,i} = [\varepsilon_{t,i-1}, 0, \dots, 0],$$

α is a $1 \times (r - 2)$ vector of constants

$$\alpha = [\alpha_2, \dots, \alpha_{r-1}].$$

The $(r + s - 1) \times 1$ matrix $B_{t,i}$ is defined as

$$B_{t,i} = [\omega^* + \alpha_0 I(\varepsilon_{t,i-1} = 0), \mathbf{0}_{1,r+s-2}]'.$$

Conditions for the stationarity and ergodicity of $g_{t,i}$ can be then derived by means of the following proposition.

Proposition 1 (Strict stationarity and ergodicity of $g_{t,i}$). Assume that A1, A2 and A3 hold, Eq. (11) will then admit a unique stationary and ergodic solution if and only if

$$\gamma(A) = \lim_{N \rightarrow \infty} \frac{1}{N} E(\log \|A_{t,i} A_{t,i-1} \dots A_{1,1}\|) \stackrel{a.s.}{=} \lim_{N \rightarrow \infty} \frac{1}{N} (\log \|A_{t,i} A_{t,i-1} \dots A_{1,1}\|) < 0, \quad (12)$$

where $\gamma(A)$ is the top Lyapunov exponent associated to the sequence $\{A_{t,i}\}$.

If Assumption A1 is replaced by the weaker Assumption A1' reported below, condition (12) will be then only sufficient for the stationarity and ergodicity of the sequence $\{g_{t,i}\}$.

Assumption A1'. The multiplicative innovation term $\varepsilon_{t,i}$ is assumed to be a non-negative, strictly stationary and ergodic process with unit mean and constant variance σ^2 .

Proof. The proof follows from the application of well known results in the theory of dynamical stochastic processes. Note that, under Assumption A1, the matrices $\{A_{t,i}, B_{t,i}\}$ are by construction i.i.d. $\forall t, i$. Also they are such that $E(\log^+ \|A_{0,0}\|)$ and $E(\log^+ \|B_{0,0}\|)$ are finite. The desired result then follows from Theorem 2.5 in [Bougerol and Picard \(1992b\)](#). Differently, in the case of stationary and ergodic errors (obtained replacing A1 by A1'), the results in [Glasserman and Yao \(1995\)](#) (Lemma 3A) or [Bougerol and Picard \(1992b\)](#) (Theorem 1.1) can be applied to prove the sufficiency of condition (12) and complete the proof. \square

In general, condition (12) cannot be exactly computed, but it can be easily approximated by computer simulation. In the simple case of a model of order (1,1), it can be conveniently rewritten as

$$E[\log(\alpha_1 \varepsilon_{t,i} + \beta_1)] < 0.$$

In addition, it can be shown (see Theorem 2.5 in [Francq and Zakoian \(2011\)](#) and associated remarks) that the following condition

$$\left(\sum_{i=1}^r \alpha_i + \sum_{j=1}^s \beta_j \right) < 1 \quad (13)$$

implies that (12) is satisfied.

In Proposition 2 we investigate the mean stationarity of $g_{t,i}$, that is we provide necessary and sufficient conditions under which $E(g_{t,i})$ is time-invariant and finite: $E(g_{t,i}) = G$, with $0 < G < \infty$.

Proposition 2 (Mean stationarity of $g_{t,i}$). *Assume that assumptions A1, A2 and A3 hold and let $\pi = P(y_{t,i} > 0)$. The process in (5) is mean stationary if and only if condition (13) is satisfied. Moreover, for a mean stationary process, the expectation of $g_{t,i}$ will be given by*

$$E(g_{t,i}) = G = \frac{\omega^* + \alpha_0(1 - \pi)}{1 - \sum_{j=1}^{r^*} (\alpha_j + \beta_j)}, \quad (14)$$

where $r^* = \max(r, s)$ and the following notational conventions have been introduced: $\alpha_j = 0$ and $\beta_k = 0$ for $j > r$ and $k > s$, respectively.

Proof. We first prove the necessary part of the condition. Note that, under Assumption A1, we have

$$E(\tilde{y}_{t,i}) = E(g_{t,i})E(\varepsilon_{t,i}) = G$$

and, under assumptions A2 and A3,

$$E[I(y_{t,i} = 0)] = E[I(\varepsilon_{t,i} = 0)].$$

It then follows that

$$\begin{aligned} G &= \omega^* + \alpha_0 E[I(\varepsilon_{t,i} = 0)] + \sum_{j=1}^r \alpha_j G + \sum_{k=1}^s \beta_k G \\ &= \omega^* + \alpha_0(1 - \pi) + \sum_{j=1}^{r^*} (\alpha_j + \beta_j) G, \end{aligned}$$

from which Eq. (14) is immediately obtained. For the sufficient part, it should be noted that condition (13) implies strict stationarity of $g_{t,i}$ in (5). Furthermore, the Markovian representation in (11) can be rewritten as

$$\mathbf{g}_{t,i} = B_{t,i} + A_{t,i}(B_{t,i-1} + A_{t,i-1}\mathbf{g}_{t,i-2}) = B_{t,i} + A_{t,i}B_{t,i-1} + A_{t,i}A_{t,i-1}\mathbf{g}_{t,i-2} \quad (15)$$

and, by n repeated substitutions, as

$$\mathbf{g}_{t,i} = B_{t,i} + \sum_{k=1}^{n-1} \prod_{j=0}^{k-1} A_{t,i-j} B_{t,i-k} + \prod_{j=0}^{n-1} A_{t,i-j} \mathbf{g}_{t,i-n}.$$

From Theorem 1.3 in [Bougerol and Picard \(1992a\)](#) and Theorem 2.4 in [Bougerol and Picard \(1992b\)](#) it follows that, as $n \rightarrow \infty$, the series $\sum_{k=1}^{n-1} \prod_{j=0}^{k-1} A_{t,i-j} B_{t,i-k}$ converges almost surely and $\prod_{j=0}^n A_{t,i-j} \rightarrow \mathbf{0}$. This implies that, for $n \rightarrow \infty$, Eq. (15) simplifies to

$$\mathbf{g}_{t,i} = B_{t,i} + \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} A_{t,i-j} B_{t,i-k}, \quad (16)$$

which is the unique strictly stationary solution to (11). Note that, under the stated assumptions, A1¹, A2 and A3, the matrices $A_{t,i}$ are positive and serially independent and the same holds for the sequence $B_{t,i}$, for any t and i . In addition, their expectations $E(A_{t,i}) = A$ and $E(B_{t,i}) = B$ will be time invariant. Finally, all the terms in the product on the RHS of (16) will be independent because each of the involved matrices depend on a single lag of $\varepsilon_{t,i-j}$ with the dates $(t, i-j)$ being distinct: $A_{t,i-j} = A(\varepsilon_{t,i-j-1})$, for $j = 0, \dots, k-1$, and $B_{t,i-k} = B(\varepsilon_{t,i-k-1})$. It follows that, taking expectations on both sides of (16) leads to

$$E(\mathbf{g}_{t,i}) = B + \sum_{k=1}^{\infty} A^k B = (I + \sum_{k=1}^{\infty} A^k) B, \quad (17)$$

where convergence of the RHS of (17) stems from the fact that the spectral radius of A is strictly less than 1 (see Theorem 2.5 in [Francq and Zakoian \(2011\)](#) and associated remarks). \square

Applying Proposition 2, Eq. (14) implies that the unit mean assumption for $g_{t,i}$ is satisfied under the additional parameter constraint on ω^*

$$\omega^* = 1 - \sum_{j=1}^{r^*} (\alpha_j + \beta_j) - \alpha_0(1 - \pi). \quad (18)$$

This expression is also useful for deriving the required positivity constraints on ω^* and $\omega^* + \alpha_0$. For example, in the case of a model of order (1,2), like the one estimated in our empirical application, the positivity constraint on ω^* can be enforced by letting

$$\frac{1 - \alpha_1 - \beta_1 - \beta_2}{1 - \pi} > \alpha_0,$$

while $(\omega^* + \alpha_0) > 0$ when

$$\frac{\alpha_1 + \beta_1 + \beta_2 - 1}{\pi} < \alpha_0,$$

where $0 < \pi < 1$. Merging the two above conditions, it follows that the constraints $\omega^* > 0$ and $(\omega^* + \alpha_0) > 0$ are simultaneously satisfied for

$$\frac{\alpha_1 + \beta_1 + \beta_2 - 1}{\pi} < \alpha_0 < \frac{1 - \alpha_1 - \beta_1 - \beta_2}{1 - \pi}. \quad (19)$$

Finally, we derive an analytical expression for $E(g_{t,i+h}|\mathcal{F}_{t,i})$, for $h > 0$, where $\mathcal{F}_{t,i}$ is the sigma-field generated by the available information until interval i of day t . First, note that, under the assumptions that $g_{t,i}$ and $\tau_{t,i}$ are strictly positive (A2 and A3), $I(y_{t,i} = 0) = I(\varepsilon_{t,i} = 0)$. Second, under the additional assumption that the errors $\varepsilon_{t,i}$ are iid (A1), it is easy to show that

$$E(\tilde{y}_{t,i+h}|\mathcal{F}_{t,i}) = E(g_{t,i+h}|\mathcal{F}_{t,i}) \quad \forall h > 0.$$

The desired conditional expectation can be then easily derived by standard calculations

$$\begin{aligned} E(g_{t,i+h}|\mathcal{F}_{t,i}) &= \omega^* + \alpha_0 E[I(y_{t,i+h-1} = 0)|\mathcal{F}_{t,i}] + \sum_{j=1}^{r^*} (\alpha_j + \beta_j) E(g_{t,i+h-j}|\mathcal{F}_{t,i}) \\ &= \omega^* + \alpha_0 E[I(\varepsilon_{t,i+h-1} = 0)] + \sum_{j=1}^{r^*} (\alpha_j + \beta_j) E(g_{t,i+h-j}|\mathcal{F}_{t,i}) \\ &= \omega^* + \alpha_0(1 - \pi) + \sum_{j=1}^{r^*} (\alpha_j + \beta_j) E(g_{t,i+h-j}|\mathcal{F}_{t,i}). \end{aligned}$$

¹Note that A1 also implies the constant zero probability assumption.

From the above formula it can be argued that, in the case of a stationary model of order (1,1), the sum $\alpha_1 + \beta_1$ can be interpreted as the *persistence* of the short term component $g_{t,i}$, defined as the speed at which, in absence of shocks, multi-step ahead predictors converge to their long-run level.

3.2. Stationarity of the seasonally adjusted volumes $y_{t,i}$

Despite its apparent complexity, the linear rolling window specification of the long-run component in Eq. (10) still allows to derive stationarity and ergodicity conditions for the seasonal adjusted volumes $y_{t,i}$. The desired result is achieved following a strategy inspired by the approach proposed in Wang and Ghysels (2015). The first step is to rewrite the model in (10) as an autoregressive model at the intra-daily frequency

$$\begin{aligned}
\tau_{t,i} &= m + \theta_d \sum_{k=1}^{K_d} \varphi_k(\omega_{1,d}, \omega_{2,d}) YD_{t,i}^{(k)} + \theta_h \sum_{k=1}^{K_d} \sum_{l=1}^H \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) YH_{t,i}^{(l,k)} \\
&= m + \sum_{k=1}^{K_d} \left[\theta_d \varphi_k(\omega_{1,d}, \omega_{2,d}) YD_{t,i}^{(k)} + \theta_h \sum_{l=1}^H \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) YH_{t,i}^{(l,k)} \right] \\
&= m + \sum_{k=1}^{K_d} \left[\theta_d \varphi_k(\omega_{1,d}, \omega_{2,d}) \sum_{j=1}^I y_{t,i-(k-1)I-j} + \theta_h \sum_{l=1}^H \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) \sum_{p=1}^{n_h} y_{t,i-(k-1)I-(l-1)n_h-p} \right] \\
&= m + \sum_{k=1}^{K_d} \left[\theta_d \sum_{l=1}^H \varphi_k(\omega_{1,d}, \omega_{2,d}) \sum_{p=1}^{n_h} y_{t,i-(k-1)I-(l-1)n_h-p} + \theta_h \sum_{l=1}^H \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) \sum_{p=1}^{n_h} y_{t,i-(k-1)I-(l-1)n_h-p} \right] \\
&= m + \sum_{k=1}^{K_d} \left[\sum_{l=1}^H \sum_{p=1}^{n_h} \theta_d \varphi_k(\omega_{1,d}, \omega_{2,d}) y_{t,i-(k-1)I-(l-1)n_h-p} + \sum_{l=1}^H \sum_{p=1}^{n_h} \theta_h \varphi_{l,k}(\omega_{1,h}, \omega_{2,h}) y_{t,i-(k-1)I-(l-1)n_h-p} \right] \\
&= m + \sum_{j=1}^{K_d I} \psi_j y_{t,i-j},
\end{aligned}$$

where

$$\psi_j = \theta_d \varphi_k(\omega_{1,d}, \omega_{2,d}) + \theta_h \varphi_{l,k}(\omega_{1,h}, \omega_{2,h})$$

for $(k-1)I + (l-1)n_h < j \leq (k-1)I + l n_h$, $k = 1, \dots, K_d$, $l = 1, \dots, H$. Letting

$$\theta^* = \sum_{j=1}^{K_d I} \psi_j,$$

we can write

$$\tau_{t,i} = m + \theta^* \sum_{j=1}^{K_d I} \frac{\psi_j}{\theta^*} y_{t,i-j} = m + \theta^* \sum_{j=1}^{K_d I} \psi_j^* y_{t,i-j},$$

where $\sum_{j=1}^{K_d I} \psi_j^* = 1$. Then note that, letting $\eta_{t,i} = g_{t,i} \varepsilon_{t,i}$, it follows that

$$y_{t,i} = \tau_{t,i} \eta_{t,i} = \left(m + \theta^* \sum_{j=1}^{K_d I} \psi_j^* y_{t,i-j} \right) \eta_{t,i}, \quad (20)$$

where $\eta_{t,i}$ is strictly stationary and ergodic under the conditions in Proposition 1. Furthermore, by some algebra, Eq. (20) can be reformulated as

$$\begin{aligned}
y_{t,i} &= \left(m + \theta^* \sum_{j=1}^{K_d I} \psi_j^* y_{t,i-j} \right) g_{t,i} \varepsilon_{t,i} \\
&= (m g_{t,i} \varepsilon_{t,i}) + \theta^* g_{t,i} \varepsilon_{t,i} \sum_{j=1}^{K_d I} \psi_j^* y_{t,i-j},
\end{aligned}$$

that can be further elaborated to give the $(K_d I \times 1)$ vector random coefficient AR(1) model

$$\mathbf{y}_{t,i} = \mathbf{B}_{t,i}^* + \mathbf{A}_{t,i}^* \mathbf{y}_{t,i-1}, \quad (21)$$

where

$$\begin{aligned} \mathbf{y}_{t,i} &= (y_{t,i}, y_{t,i-1}, \dots, y_{t,i-K_d I+1})', \\ \mathbf{B}_{t,i}^* &= (m g_{t,i} \varepsilon_{t,i}, 0, \dots, 0)', \\ \mathbf{A}_{t,i}^* &= \begin{pmatrix} \boldsymbol{\Psi}_{t,i}' & \theta^* g_{t,i} \varepsilon_{t,i} \psi_{K_d I} \\ \mathbf{I}_{K_d I-1} & \mathbf{0}_{K_d I-1,1} \end{pmatrix}, \end{aligned}$$

with

$$\boldsymbol{\Psi}_{t,i} = (\theta^* g_{t,i} \varepsilon_{t,i} \psi_1, \dots, \theta^* g_{t,i} \varepsilon_{t,i} \psi_{K_d I-1})'.$$

Conditions for stationarity and ergodicity of $\mathbf{y}_{t,i}$ are then established by the following proposition.

Proposition 3 (Strict stationarity and ergodicity of $\mathbf{y}_{t,i}$). *Assume that A1, A2, A3 and the condition in Eq. (13) hold. Eq. (21) will admit a unique strictly stationary and ergodic solution iff*

$$\gamma(A^*) = \lim_{N \rightarrow \infty} \frac{1}{N} E(\log \|\mathbf{A}_{t,i}^* \mathbf{A}_{t,i-1}^* \dots \mathbf{A}_{1,1}^*\|) \stackrel{a.s.}{=} \lim_{N \rightarrow \infty} \frac{1}{N} (\log \|\mathbf{A}_{t,i}^* \mathbf{A}_{t,i-1}^* \dots \mathbf{A}_{1,1}^*\|) < 0. \quad (22)$$

Proof. The proof closely follows that of Proposition 3.3 of Wang and Ghysels (2015). To start, first note that we can rewrite $\mathbf{A}_{t,i}^*$ in (21) as follows

$$\mathbf{A}_{t,i}^* = \theta^* g_{t,i} \varepsilon_{t,i} \mathbf{H} \mathbf{D} + \mathbf{F},$$

where \mathbf{H} and \mathbf{F} are square matrices of dimension $(K_d I)$ such that

$$\mathbf{H} = \begin{pmatrix} \mathbf{1}_{1,K_d I} \\ \mathbf{0}_{K_d I-1,K_d I} \end{pmatrix} \quad \mathbf{F} = \begin{pmatrix} \mathbf{0}_{1,K_d I-1} & 0 \\ \mathbf{I}_{K_d I-1} & \mathbf{0}_{K_d I-1,1} \end{pmatrix}$$

and $\mathbf{D} = \text{Diag}(\boldsymbol{\psi}^*)$ is a diagonal matrix with main diagonal given by $\boldsymbol{\psi}^* = (\psi_1^*, \psi_2^*, \dots, \psi_{K_d I}^*)'$. Without loss of generality, we set $t = 0$ and $i = 0$ and note that $1 \leq \|\mathbf{A}_{0,0}^*\| \leq \theta^* g_{0,0} \varepsilon_{0,0} \|\mathbf{H} \mathbf{D}\| + \|\mathbf{F}\| \leq \theta^* g_{0,0} \varepsilon_{0,0} + 1$, using the sub-multiplicative nature of the induced matrix norm and the fact that $\|\mathbf{H} \mathbf{D}\| = 1$, given that $0 \leq \psi_j^* < 1$ (for $j = 1, \dots, K_d I$). Therefore, $E(\log(\|\mathbf{A}_{0,0}^*\|)) \leq E(\log(\|\theta^* g_{0,0} \varepsilon_{0,0} + 1\|)) \leq E(\log(\theta^* g_{0,0} \varepsilon_{0,0} + 1)) < \infty$, using the result $\log(1+x) \leq x$, $\forall x > -1$ (Love, 1980). Sufficiency follows from Theorem 3.1 in Glasserman and Yao (1995). To prove the necessity, we use a strategy similar to that adopted in the proof of Theorem 1.3 of Bougerol and Picard (1992a).

Suppose that Eq. (21) admits a unique strictly stationary and ergodic solution and let ${}_n S_{t,i} = \prod_{j=0}^{n-1} \mathbf{A}_{t,i-j}^*$, for $n > 0$, and ${}_0 S_{t,i} = \mathbf{I}$, for $n = 0$. It is easy to show, by repeated substitutions, that

$$\mathbf{y}_{0,0} = {}_k S_{0,0} \mathbf{y}_{0,-k} + \sum_{n=0}^{k-1} {}_n S_{0,0} \mathbf{B}_{0,-n}^*. \quad (23)$$

From the positivity of the components on both sides of (23) it follows that $\sum_{n=0}^{k-1} {}_n S_{0,0} \mathbf{B}_{0,-n}^* \leq \mathbf{y}_{0,0}$, a.s. for any k . Convergence of this summation, in turn, implies that a.s. $\lim_{n \rightarrow \infty} {}_n S_{0,0} \mathbf{B}_{0,-n}^* = \mathbf{0}$. Letting $\{e_1, e_2, \dots, e_{K_d I}\}$ be the canonical basis of $\mathbb{R}^{K_d I}$, this also implies $\lim_{n \rightarrow \infty} \tilde{\mathbf{y}}_{0,-n} {}_n S_{0,0} e_1 = \mathbf{0}$. Moreover, for $i = 1, \dots, K_d I - 1$, the following relation holds

$${}_n S_{0,0} e_i = {}_{n-1} S_{0,0} (\theta^* \tilde{\mathbf{y}}_{0,-n+1} \mathbf{H} \mathbf{D} + \mathbf{F}) e_i = (\theta^* \psi_i^* \tilde{\mathbf{y}}_{0,-n+1}) {}_{n-1} S_{0,0} e_1 + {}_{n-1} S_{0,0} e_{i+1}, \quad (24)$$

while, for $i = K_d I$, we have

$${}_n S_{0,0} e_{K_d I} = (\theta^* \psi_{K_d I}^* \tilde{\mathbf{y}}_{0,-n+1}) {}_{n-1} S_{0,0} e_1. \quad (25)$$

Eq. (25) and Eq. (24) can be recursively applied to prove that $\lim_{n \rightarrow \infty} {}_n S_{0,0} e_i = \mathbf{0}$, a.s. for $i = 1, \dots, K_d I$. It follows that $\lim_{n \rightarrow \infty} {}_n S_{0,0} = \mathbf{0}$, a.s.. Finally, Lemma 3.4 in Bougerol and Picard (1992b) applies to prove that $\gamma(A^*) < 0$. This completes the proof. \square

4. Inference

4.1. The Zero-Augmented Generalized F distribution

Multiplicative Error Models are usually estimated by QMLE, assuming that the density of the innovation term follows a Gamma distribution or one of its generalisations or special cases. However, when dealing with non-liquid assets, these distributions cannot account for a point mass at zero, as the corresponding log-likelihood functions exclude zero realizations, with the exception of the Exponential distribution, that can be obtained as a particular case of the Gamma. The specification of an Exponential distribution for $\varepsilon_{t,i}$ is a natural choice, as this distribution can be seen as the counterpart of the Normal distribution for positive-valued random variables. Under the assumption of correct specification of the conditional mean function, maximisation of the Exponential quasi log-likelihood function leads to consistent and asymptotically normal estimates of the conditional mean parameters. A formal derivation of this result, relying on the work of [Lee and Hansen \(1994\)](#) for GARCH models, can be found in [Engle and Russell \(1998\)](#) or [Engle \(2002\)](#).

However, the continuous nature of the Exponential distribution implies that the proportion of zeros must be trivial to avoid misspecification at the lower boundary of the support. It follows that, in the presence of zero observations, the Generalized Method of Moments (GMM) can be a valid alternative estimation strategy, as discussed in [Brownlees et al. \(2011\)](#), since it does not require the adoption of a specific density function for the innovation term.

In general, both Exponential-QML and GMM can yield consistent estimates of conditional mean parameters, but these become quite inefficient in the presence of a high proportion of zeros. To address this problem [Hautsch et al. \(2014\)](#) proposed an alternative estimation strategy based on the introduction of what they call *Zero-Augmented Generalized F* (ZAF) distribution. Their results provide evidence that, in the presence of a non-trivial proportion of zero outcomes, MLE based on the ZAF distribution allows to overcome the potential inconsistency of the standard QMLE and, in any case, to obtain substantial efficiency gains over the latter.

Following [Hautsch et al. \(2014\)](#), consider a non-negative random variable Z , assigning a discrete probability mass to exact zero values as follows

$$\pi = P(Z > 0), \quad (1 - \pi) = P(Z = 0), \quad (26)$$

with $0 \leq \pi \leq 1$. We will say that the variable Z follows a ZAF distribution if, conditionally on $Z > 0$, it is distributed as a Generalized F distribution with density function

$$g(z; \zeta) = \frac{a z^{ab-1} [c + (z/\nu)^a]^{-(c-b)} c^c}{\nu^{ab} \mathcal{B}(b, c)}, \quad (27)$$

where $\zeta = (a, b, c, \nu)'$, $\mathcal{B}(\cdot, \cdot)$ is the Beta function with $\mathcal{B}(b, c) = [\Gamma(b)\Gamma(c)]/\Gamma(b+c)$, $a > 0$, $b > 0$, $c > 0$ and $\nu > 0$. The Generalized F distribution is based on a scale parameter ν and three shape parameters a , b and c , thus it is very flexible, nesting different error distributions, such as the Weibull for $b = 1$ and $c \rightarrow \infty$, the Generalized Gamma for $c \rightarrow \infty$ and the Log-Logistic for $b = 1$ and $c = 1$ ([Hautsch, 2003](#)).

The overall ZAF distribution is semi-continuous with density function given by

$$f_Z(z) = (1 - \pi)\eta(z) + \pi g(z)I_{(z>0)}, \quad (28)$$

where $I_{(z>0)}$ denotes an indicator function taking the value 1 for $z > 0$ and 0 elsewhere. It can be easily noted that the ZAF density reduces to the Generalized F for $\pi = 1$.

The moments of the ZAF distribution are given by

$$E[Z^r] = \pi E[Z^r | Z > 0] + (1 - \pi)E[Z^r | Z = 0] = \pi \nu^r c^{r/a} \frac{\Gamma(b + r/a)\Gamma(c - r/a)}{\Gamma(b)\Gamma(c)}, \quad r < ac. \quad (29)$$

In order to use the ZAF distribution as a probabilistic model for the error term $\varepsilon_{t,i}$ in the MEM structure in (1), it is necessary to set $\nu = (\pi\xi)^{-1}$ to ensure that the unit mean assumption for $\varepsilon_{t,i}$ is fulfilled and

$$\xi = c^{1/a} [\Gamma(b + 1/a)\Gamma(c - 1/a)] [\Gamma(b)\Gamma(c)]^{-1}. \quad (30)$$

4.2. The Dynamic Zero-Augmented Generalized F distribution

The presence of zero volumes is very common in high frequency trading and, thus, their behaviour needs to be modelled. As discussed above, this can be done using the ZAF distribution. However, there could be cases in which it is not reasonable to assume that the value of the trading probability is constant over time. In particular, this is likely to occur if the time interval under investigation is sufficiently long and characterised by the alternance of periods featuring remarkably different volatilities and trading intensities. To account for the presence of time-varying zero probabilities, [Hautsch et al. \(2014\)](#) proposed a dynamic version of the ZAF distribution, that is the *Dynamic Zero-Augmented Generalized F* (DZAF). Assuming a DZAF distribution for $\varepsilon_{t,i}$ is equivalent to assume that the trading probability, which is the probability of observing non-zero volumes, is time-varying. Namely,

$$\pi_{t,i} = P(\varepsilon_{t,i} > 0 | \mathcal{F}_{t,i-1}) = P(\varepsilon_{t,i} | \mathcal{H}_{t,i-1}) = \pi(\mathcal{H}_{t,i-1}, \boldsymbol{\theta}_\pi),$$

where $\mathcal{H}_{t,i-1} \subset \mathcal{F}_{t,i-1}$ and $\boldsymbol{\theta}_\pi$ the parameter vector characterising $\pi_{t,i}$.

Now, let $\mathcal{I}_{t,i}$ be a binary trade indicator taking value 1 if $\varepsilon_{t,i} > 0$ and 0 otherwise. The time-varying probability $\pi_{t,i}$ can be modelled by means of a logistic function of the kind

$$\pi_{t,i} = \frac{\exp(h_{t,i})}{1 + \exp(h_{t,i})}, \quad (31)$$

where $h_{t,i}$ is assumed to follow the Autoregressive Conditional Multinomial (ACM) specification proposed by [Russell and Engle \(2005\)](#)

$$h_{t,i} = \varpi + \delta_1 s_{t,i-1} + \gamma_1 h_{t,i-1}, \quad (32)$$

with $s_{t,i}$ being a standardised trade indicator

$$s_{t,i} = \frac{\mathcal{I}_{t,i} - \pi_{t,i}}{\sqrt{\pi_{t,i}(1 - \pi_{t,i})}}. \quad (33)$$

Since $\{s_{t,i}\}$ is a martingale difference with zero mean and unit variance, it follows that $\{h_{t,i}\}$ is an ARMA process with a weak white noise error term, resulting stationary if $|\gamma_1| < 1$.

The main consequence related to the time-varying zero probability assumption is that the error terms $\varepsilon_{t,i}$ lose the i.i.d. property since, conditionally on the information set $\mathcal{H}_{t,i-1}$, they become independently but not identically distributed. In this case, the conditional density of $\varepsilon_{t,i} | \mathcal{H}_{t,i-1}$ is then given by

$$f_\varepsilon(\varepsilon_{t,i} | \mathcal{H}_{t,i-1}) = (1 - \pi_{t,i}) \eta(\varepsilon_{t,i}) + \pi_{t,i} g_\varepsilon(\varepsilon_{t,i}) I_{(\varepsilon_{t,i} > 0)}. \quad (34)$$

It is worth remarking that, under the assumption of DZAF errors, Assumption A1 is not fulfilled implying that Propositions 1-3 cannot be immediately extended to the models with DZAF errors illustrated in this section.

4.3. Two stage estimation procedure

The estimation of the H-MIDAS-CMEM model is performed in two stages. In the first stage, the parameters of the Fourier Flexible Form specified in (3) for the seasonal factors ϕ_i are estimated by an OLS regression of the raw volumes $x_{t,i}$ on the regressors on the RHS of Eq. (3). The seasonal adjusted volumes $\hat{y}_{t,i}$ are estimated as $x_{t,i} / \hat{\phi}_i$, where $\hat{\phi}_i$ is the estimated seasonal factor for the i -th intra-daily period. In the second stage, conditional on the estimated seasonal factors, the unknown parameters in $g_{t,i}$ and $\tau_{t,i}$ are estimated by maximising likelihood functions based on the assumptions of ZAF or DZAF errors.

We now move to deriving the second-stage likelihood functions under the ZAF and DZAF assumptions on the conditional distribution of $\varepsilon_{t,i}$. In the ZAF case, the second stage log-likelihood function for $\hat{y}_{t,i} = x_{t,i} / \hat{\phi}_i$, based on the density in (28), is given by

$$\begin{aligned} \mathcal{L}(\mathbf{y}; \boldsymbol{\theta}, \pi) = & n_z \log(1 - \pi) + n_{nz} \log \pi + \sum_{t,i \in \mathcal{J}_{nz}} \left\{ \log a + (ab - 1) \log \left(\frac{\hat{y}_{t,i}}{\tau_{t,i} g_{t,i}} \right) + c \log c \right. \\ & \left. - (c + b) \log \left[c + \left(\frac{\pi \xi \hat{y}_{t,i}}{\tau_{t,i} g_{t,i}} \right)^a \right] - \log(\tau_{t,i} g_{t,i}) - \log \mathcal{B}(b, c) + ab \log(\pi \xi) \right\}, \end{aligned} \quad (35)$$

where \mathcal{J}_{n_z} denotes the set of all observations different from zero, while n_z and n_{n_z} are the number of zero and non-zero observations respectively, with $\boldsymbol{\theta} = (\boldsymbol{\theta}'_g, \boldsymbol{\theta}'_\tau, a, b, c)'$.

Similarly, in the DZAF case, the log-likelihood for $\hat{y}_{t,i}$, based on the density function in (34), can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\theta}_\pi) = & \sum_{t,i} I_{t,i} \log(\pi_{t,i}) + (1 - I_{t,i}) \log(1 - \pi_{t,i}) + \sum_{t,i \in \mathcal{J}_{n_z}} \left\{ \log a + (ab - 1) \log \left(\frac{\hat{y}_{t,i}}{\tau_{t,i} g_{t,i}} \right) + c \log c \right. \\ & \left. - (c + b) \log \left[c + \left(\frac{\pi_{t,i} \xi \hat{y}_{t,i}}{\tau_{t,i} g_{t,i}} \right)^a \right] - \log(\tau_{t,i} g_{t,i}) - \log \mathcal{B}(b, c) + ab \log(\pi_{t,i} \xi) \right\}, \end{aligned} \quad (36)$$

where ξ is defined as in (30).

Since a two-stage estimation approach is used, when computing the standard errors of the second stage estimates obtained through the maximisation of (35) and (36), the uncertainty arising from the first stage should be accounted for.

This can be conveniently done representing the overall estimation problem as a two stage just-identified GMM estimator, as in Engle and Sokalska (2012). The model parameter vector can be partitioned as $\boldsymbol{\lambda}' = (\mathbf{a}', \boldsymbol{\theta}')$, where $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_\tau, \boldsymbol{\theta}'_\pi)$ and \mathbf{a} is the vector of first stage parameters estimated in the regression model in (3). The moment conditions for the estimation of \mathbf{a} are recovered from the normal equations of the OLS estimator of the regression model used for the first stage estimation of seasonal coefficients

$$u_j^{(N)} = \frac{1}{N} \sum_{t=1}^T \sum_{i=1}^I u_{ti,j} = 0,$$

with

$$u_{ti,j} = \frac{\partial}{\partial a_j} (x_{t,i} - \phi_i)^2, \quad j = 1, \dots, k_1,$$

where ϕ_i is defined as in (3) and k_1 is the number of first stage parameters in \mathbf{a} .

In a similar fashion, the second stage moment conditions for the estimation of the elements of $\boldsymbol{\theta}$, are given by the score equations of the second stage log-likelihood

$$s_\ell^{(N)} = \frac{1}{N} \sum_{t=1}^T \sum_{i=1}^I s_{ti,\ell} = 0,$$

with

$$s_{ti,\ell} = \frac{\partial}{\partial \theta_\ell} \mathcal{L}_{ti}, \quad \ell = 1, \dots, k_2$$

where $\mathcal{L}_{ti} = \mathcal{L}(y_{t,i}; \boldsymbol{\theta} | \mathbf{a})$ is the contribution of observation i of day t to the overall likelihood and k_2 is the number of second stage parameters in $\boldsymbol{\theta}$. Letting $\mathbf{u}(\mathbf{a}) = (u_1^{(N)}, \dots, u_{k_1}^{(N)})'$ and $\mathbf{s}(\boldsymbol{\theta}, \mathbf{a}) = (s_1^{(N)}, \dots, s_{k_2}^{(N)})'$, the overall vector of averaged moment conditions is denoted by

$$\mathbf{w}(\boldsymbol{\theta}, \mathbf{a}) = (\mathbf{u}(\mathbf{a})', \mathbf{s}(\boldsymbol{\theta}, \mathbf{a})')'.$$

Estimation is then performed in two steps. First, we solve $\mathbf{u}(\mathbf{a}) = \mathbf{0}$ with respect to \mathbf{a} and obtain $\hat{\mathbf{a}}$, which is the OLS estimator of the first stage parameters. Second, conditional on first stage estimates, we solve $\mathbf{s}(\boldsymbol{\theta}, \hat{\mathbf{a}}) = \mathbf{0}$ with respect to $\boldsymbol{\theta}$ and obtain the estimator of second stage parameters $\hat{\boldsymbol{\theta}}$. By Theorem 6.1 in Newey and McFadden (1994), if the usual conditions for the consistency of the OLS estimator of \mathbf{a} ($\hat{\mathbf{a}}$) hold, assuming that the $y_{t,i}$ are strictly stationary and ergodic, the consistency of $\hat{\boldsymbol{\theta}}$ will follow under standard regularity conditions and the overall GMM estimator $\hat{\boldsymbol{\lambda}}$ will be asymptotically normal

$$\sqrt{N}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) \xrightarrow{d} \mathcal{N}_{k_1+k_2}(\mathbf{0}, W^{-1} \Omega (W^{-1})'),$$

where λ_0 denotes the vector of *true* unknown model parameters,

$$W = E \left[\frac{\partial \mathbf{w}_{ti}(\lambda)}{\partial \lambda'} \right]$$

and

$$\Omega = E [\mathbf{w}_{ti}(\lambda) \mathbf{w}_{ti}(\lambda)'],$$

where $\mathbf{w}_{ti} = (\mathbf{u}_{ti}', \mathbf{s}_{ti}')'$, with $\mathbf{u}_{ti} = (u_{ti,1}, \dots, u_{ti,k_1})'$ and $\mathbf{s}_{ti} = (s_{ti,1}, \dots, s_{ti,k_2})'$. Furthermore, the matrix W has the following block structure

$$W = E \begin{bmatrix} \frac{\partial \mathbf{u}_{ti}}{\partial \mathbf{a}'} & \mathbf{0} \\ \frac{\partial \mathbf{s}_{ti}}{\partial \mathbf{a}'} & \frac{\partial \mathbf{s}_{ti}}{\partial \boldsymbol{\theta}'} \end{bmatrix}.$$

As in Hansen (1982), the matrices W and Ω can be consistently estimated replacing the expectations with sample means evaluated at $\hat{\lambda}$ and numerically evaluating the derivatives involved in W . In this way, it is possible to obtain consistent estimates of the standard errors of $\hat{\boldsymbol{\theta}}$ that take into account uncertainty in the estimation of \mathbf{a} .

Finally, it is worth dedicating some considerations to the identification of model parameters. The multiplicative component structure makes our model naturally prone to the rise of potential identification problems. Namely, without imposing adequate parametric constraints, the scale parameter of the error distribution ν ($\nu_{t,i}$ in the DZAF case), (ω^*, α_j) ($j \geq 0$), in the specification of $g_{t,i}$, and (m, θ_d, θ_h) , in the specification of $\tau_{t,i}$, would not be simultaneously identifiable. In order to overcome this problem, in addition to the positivity constraints in A2 and A3, we impose that

1. $\nu = (\pi\xi)^{-1}$ ($\nu_{t,i} = (\pi_{t,i}\xi)^{-1}$, $\forall t, i$, in the DZAF case), ensuring $E(\varepsilon_{t,i}) = 1$;
2. under the assumption of mean stationarity of $g_{t,i}$ in (13), ω^* is parameterised as in (18) ensuring $E(g_{t,i}) = 1$.

Additional threats to identification could come from the long-run component $\tau_{t,i}$. While joint identifiability of the slope coefficients (θ_d, θ_h) and weighting function parameters (ω_d, ω_h) is guaranteed by assuming that the MIDAS weights sum up to 1 (see Eq. (7)), it should be mentioned that the parameters of the weighting function in the MIDAS filters in $\tau_{t,i}$ become unidentified when the corresponding slope parameters $(\theta_d$ and $\theta_h)$ are equal to 0 that is however ruled out by Assumption A3. It should be further noted that this problem is not specific to our modelling approach and it is well known in the literature on MIDAS models. Its implications for testing hypotheses of the kind $\theta_d = 0$ or $\theta_h = 0$ are discussed in Ghysels et al. (2006), who also present alternative approaches for adjusting the test statistic and associated p-values.

In general, the complex structure of the proposed model makes the derivation of a formal proof of identifiability a complicated task. However, it should be remarked that, for all the series considered, the likelihood optimisation procedure always returns well defined solutions associated to positive definite estimated information matrices, thus providing empirical support to the local identifiability of the fitted models (Rothenberg, 1971). In addition, in order to safeguard against the presence of multiple local maxima, we performed estimation starting from different initial conditions, without experiencing dependence of the final estimates on the chosen set of initial parameter values.

5. Empirical application

5.1. Data description

The high frequency trading volume data used in our analysis refer to the stocks Beiersdorf (BEI), a personal-care company, Continental (CON), a manufacturing company specialised in tyres, brake systems and other vehicles parts, Deutsche Telekom (DTE), one of the most important telecommunications company in Europe, GEA Group (G1A), one of the largest suppliers of equipment and process technology, mainly for the food industry, Salzgitter (SZG), a leading steel manufacturing company, and Volkswagen (VOW), a multinational manufacturers of automobiles and commercial vehicles. These assets are all traded on the Xetra Market in the German Stock Exchange. The raw tick-by-tick volume data have been filtered employing the procedure described in Brownlees and Gallo (2006) considering regular trading hours from 9:00 am to 5:30 pm. The filtered volumes have been aggregated over 10-minutes intervals, which means 51 observations per day. The empirical analysis covers the period between 2 January 2009 and 27

December 2012 including 1017 trading days and 51867 intra-daily observations for each stock. The time plots of the six time series of 10-minute trading volumes have been reported in the online Empirical Appendix (Figure 8).

Figure 1 displays the intraday seasonal component estimated via the Fourier Flexible Form defined in Eq. (3). As expected, the average trading intensity varies across trading hours giving rise to a typical U-shaped pattern. This is consistent with the well known stylised fact by which time between trades tends to be shorter near the open and the close of the trading day than in the middle of the day, as documented in Engle and Russell (1998).

Figure 1: Estimated intra-daily periodic component for regular trading hours: 9:00 am – 5:30 pm

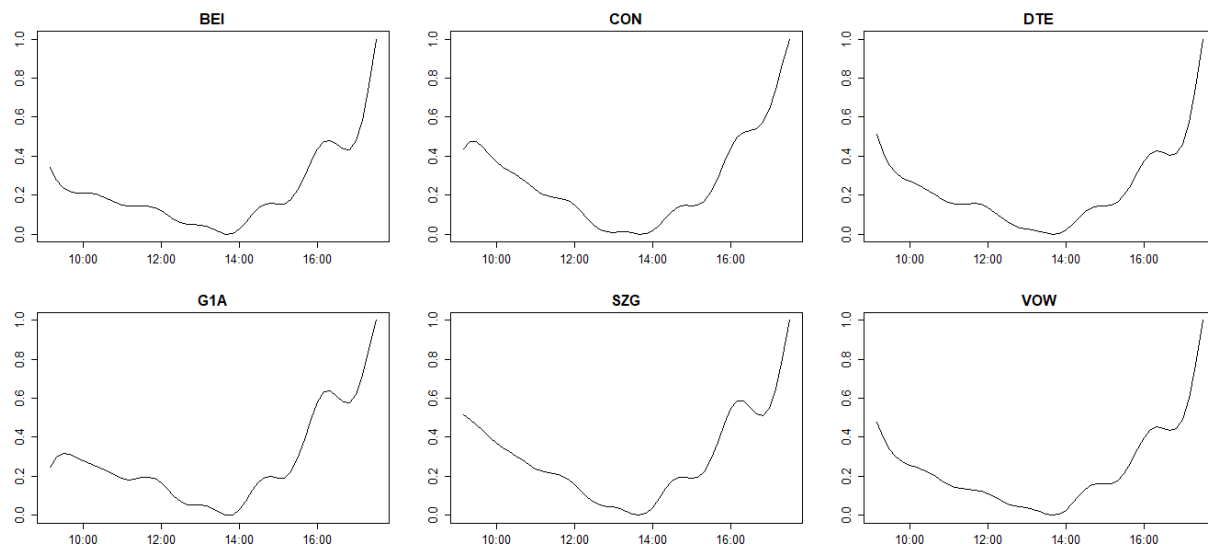


Table 1: Summary Statistics

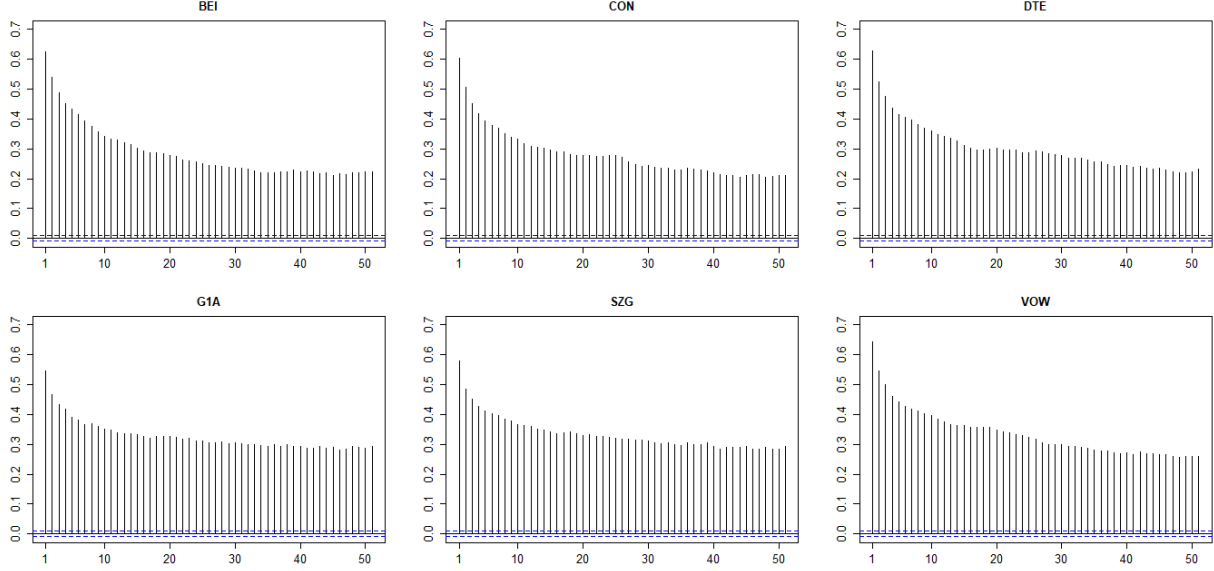
Ticker	Zero%	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.	$\hat{\rho}_1$	$\hat{\rho}_{51}$
BEI	0.064	0.000	2144	3818	5249	6650	135800	5176.852	0.663	0.349
CON	0.792	0.000	1590	3049	4143	5380	117600	4149.137	0.647	0.306
DTE	0.014	0.000	64840	103200	132400	165100	2088000	108350.631	0.690	0.407
G1A	0.251	0.000	2734	5251	7732	9878	134100	8082.502	0.605	0.398
SZG	0.102	0.000	1468	2758	3831	5022	87830	3570.694	0.643	0.406
VOW	0.390	0.000	731	1488	2282	2853	55680	2690.027	0.691	0.374
BEI	0.064	0.000	0.461	0.763	1.000	1.232	29.630	0.953	0.627	0.223
CON	0.792	0.000	0.425	0.770	1.000	1.281	22.360	0.957	0.605	0.211
DTE	0.014	0.000	0.551	0.820	1.000	1.221	22.930	0.754	0.629	0.233
G1A	0.251	0.000	0.398	0.709	1.000	1.250	16.440	1.004	0.546	0.292
SZG	0.102	0.000	0.428	0.749	1.000	1.297	26.480	0.882	0.579	0.293
VOW	0.390	0.000	0.356	0.687	1.000	1.258	34.360	1.123	0.644	0.259

Summary statistics of 10 minute raw (top panel) and seasonally adjusted (bottom panel) trading volumes. Zero%: Percentage of zero observations; Min.: Minimum; 1st Qu.: First Quartile; Median; Mean; 3rd Qu.: Third Quartile; Max.: Maximum; Std. Dev.: Standard Deviation; $\hat{\rho}_1$: Autocorrelation at lag 1; $\hat{\rho}_{51}$: Autocorrelation at the lag 51 (1 day). Sample period: 02 Jan 2009 - 27 Dec 2012.

Descriptive statistics of the raw (top-panel) and seasonally adjusted (bottom-panel) 10-minute trading volumes are shown in Table 1. An important feature of the data is the presence of non-trading intervals leading to zero volumes with frequency ranging from 0.0135% to 0.7924%, for DTE and CON respectively. Dividing the data by the estimated periodic component, trading volumes are rescaled to have unit mean.

The last two columns of [Table 1](#) report the values of the sample autocorrelation at the lags 1 and 51 (1 day). The seasonal adjusted volumes, as it is also evident from [Figure 2](#), are still characterised by a highly persistent autocorrelation structure featuring a decay pattern apparently much slower than what implied by the assumption of exponential decay.

Figure 2: Autocorrelation function of seasonally adjusted volumes



5.2. In sample estimation results and model diagnostics

In this section we fit a H-MIDAS-CMEM to the time series of trading volumes for the stocks BEI, CON, DTE, G1A, SZG and VOW, comparing its in-sample performances to those of a number of alternative specifications. All the models are based on seasonally adjusted volumes constructed from the same estimated ϕ_i component with $Q = 2$ and $P = 6$, identified according to the BIC. This result is in line with the findings in [Andersen et al. \(2000\)](#). Also, for the sake of parsimony, following a common practice in the literature on MIDAS models, we impose the following constraints on the parameters of the Beta weighting function in Eq. (7): $\omega_{1,d} = \omega_{1,h} = 1$; $\omega_{2,d} > 1$ and $\omega_{2,h} > 1$. This choice returns weight sequences that are monotonically decreasing over the lags.

As benchmarks we have considered the standard MEM, the CMEM of [Brownlees et al. \(2011\)](#), with

$$\tau_t = m + \alpha_{1,d} \bar{y}_{t-1} + \beta_{1,d} \tau_{t-1}$$

and a standard Corsi-style HAR-CMEM specification, with

$$\tau_t = m + \beta_{1,d} \bar{y}_{t-1} + \beta_{1,w} \bar{y}_{t-1:t-5} + \beta_{1,m} \bar{y}_{t-1:t-20},$$

where $\bar{y}_t = I^{-1} \sum_{i=1}^I y_{t,i}$, $\bar{y}_{t-1:t-5} = \sum_{j=1}^5 \bar{y}_{t-j}$ and $\bar{y}_{t-1:t-20} = \sum_{j=1}^{20} \bar{y}_{t-j}$ are the daily, weekly and monthly average volumes, respectively. In addition, we consider the *homogeneous* version of the H-MIDAS-CMEM, denoted as MIDAS-CMEM, where the long term component is specified as in (6) and only includes a daily MIDAS filter.

The model parameters have been estimated using two different specifications of the second stage log-likelihood based on the ZAF and DZAF distributions, respectively. For the specifications with ZAF errors, the estimated model configuration has been selected to minimise the BIC over a grid of values ranging from 1 to 2, for the orders r and s of the short term component $g_{t,i}$, and from 200 to 500, with step equal to 20, for the number of daily MIDAS lags K_d . [Table 2](#) reports the selected number of lags for models featuring a MIDAS-type long-run component. In the DZAF case, in order to more easily assess the impact of assuming a time-varying probability, we keep the same model structure identified for models with ZAF errors.

Table 2: Values of K_d minimising the BIC criterion for the MIDAS-CMEM and H-MIDAS-CMEM fitted to the time series of trading volumes on six German stocks traded on the Xetra Market. K_d denotes the maximum daily lag covered by the filter.

	BEI	CON	DTE	G1A	SZG	VOW
MIDAS	240	440	300	320	380	260
H-MIDAS	240	460	320	300	380	240

Table 3 reports the parameter estimates obtained under the assumption of constant trading probability (ZAF). Coefficients that are not significant at the usual 5% significance level are reported in bold. Similarly, parameter estimates obtained under the assumption of time-varying trading probability (DZAF) are reported in Table 4. In both cases the estimates are based on the full available sample 2009-2012. The associated standard errors can be found in the online Empirical Appendix in tables 11 and 12, respectively. It is worth noting that standard errors for ω^* are missing due to the fact that this parameter is estimated through expectation targeting.

In general, we find that all the parameters are significantly different from zero with a few exceptions. Namely, the dummy coefficient of the short term component α_0 is always significant only for the DTE series while, for CON, we manage to reject the null $\alpha_0 = 0$ only for the H-MIDAS-CMEM. For the other series the estimated α_0 is never significantly different from zero. The trend intercept m is not significant for the H-MIDAS-CMEM fitted to CON and the CMEM fitted to G1A.

Focusing on the short term component parameters, we note that the α_1 and β_1 coefficients tend to assume values remarkably close across different assets and models. Differently, probably due to the inclusion of the additional higher frequency MIDAS filter, the value of the estimated β_2 tends to take lower values for the H-MIDAS-CMEM model. When analysing the estimates obtained for the trend parameters, three main facts arise. First, the values of ω_d and ω_h , the shape parameters of the Beta weighting function, are substantially varying across assets. Second, the value of the estimated ω_d substantially decreases as the hourly filter is added to the model, that is when we move from the MIDAS-MEM to the H-MIDAS-CMEM. This means that the introduction of the hourly filter has the effect of increasing the memory of the daily one. Third, in the H-MIDAS-CMEM, as expected, we always find $\omega_d < \omega_h$, implying that the weights of the daily filter are more slowly decaying than those involved in the hourly filter. Finally, looking at the fitted error distributions, the estimates of the static ZAF parameters appear quite stable across different models, while their variation across assets is more pronounced. Furthermore, it is worth noting that the estimated parameter π of the ZAF distribution is very close to the empirical frequency of non-zero outcomes $\hat{\pi} = N^{-1} \sum_{i=1}^N I_{(y_{t,i} > 0)}$.

Similar considerations apply to the estimates based on the Dynamic ZAF distribution in Table 4. In general, the estimates of the intra-daily, trend and distribution parameters a , b and c are very close to those in Table 3 based on the static ZAF distribution. Also, it is important to remark that the estimates of the parameters of the time-varying probability $\pi_{t,i}$ are always significant with the exception of the intercept ϖ for the MIDAS-CMEM fitted to the stock G1A. Moreover, the estimated value of γ_1 is always < 1 in module, suggesting stationarity of the fitted ACM model for $\{h_{t,i}\}$.

Table 5 reports the BIC and log-likelihood values of the fitted models. The BIC values recorded for the simple MIDAS-CMEM and H-MIDAS-CMEM are always remarkably lower than those of the competing models, with the H-MIDAS-CMEM returning the lowest values. The CMEM and HAR-CMEM are characterised by similar performances, while the standard MEM seems to be the weakest competitor in terms of BIC values. These findings enhance the empirical evidence in favour of the hypothesis that trading volumes tend to cluster around a time-varying lower frequency component, thus providing support for the use of component models. Table 5 also allows to compare the ZAF and DZAF specifications in terms of in-sample fit. Namely, this is explicitly done in the last two columns, reporting the gains in terms of log-likelihood and BIC values obtained using the DZAF and ZAF distributions, respectively. Despite the relatively low number of zero-volumes, the use of the DZAF produces a noticeable improvement in terms of both log-likelihood and BIC. Positive values of $\mathcal{L}^- = \mathcal{L}(DZAF) - \mathcal{L}(ZAF)$ indicate that the log-likelihood provided by the DZAF distribution is higher than the one obtained using the ZAF. On the other hand, the negative sign of $BIC^- = BIC(DZAF) - BIC(ZAF)$ indicates that the BIC is lower for the DZAF-models rather than for the ZAF-models.

Table 3: In sample parameter estimates based on the ZAF distribution

Ticker	Model	Intra-daily parameters						Trend parameters										ZAF parameters				
		ω^*	α_1	α_0	β_1	β_2	m	$\alpha_{1,d}$	$\beta_{1,d}$	$\beta_{1,w}$	$\beta_{1,m}$	θ_d	θ_h	ω_d	ω_h	a	b	c	π^\dagger			
BEI	MEM	0.042	0.304	-0.005	0.345	0.309	-	-	-	-	-	-	-	-	-	3.189	0.738	1.185	99.93			
	CMEM	0.087	0.305	-0.029	0.350	0.257	0.094	0.220	0.675	-	-	-	-	-	-	3.184	0.740	1.213	99.92			
	HAR	0.081	0.304	0.078	0.312	0.302	0.287	-	0.205	0.325	0.151	-	-	-	-	3.406	0.676	1.099	99.93			
	MIDAS	0.151	0.300	-0.061	0.336	0.213	0.178	-	-	-	-	0.016	-	-	-	3.241	0.726	1.185	99.92			
CON	H-MIDAS	0.236	0.290	-0.095	0.325	0.149	0.126	-	-	-	-	0.006	0.032	130.721	732.985	3.198	0.738	1.213	99.92			
	MEM	0.066	0.341	1.284	0.318	0.275	-	-	-	-	-	-	-	-	-	3.093	0.951	1.368	99.96			
	CMEM	0.096	0.361	1.092	0.306	0.237	0.064	0.221	0.726	-	-	-	-	-	-	3.088	0.973	1.321	99.96			
	HAR	0.090	0.353	0.265	0.359	0.198	0.136	-	0.195	0.314	0.403	-	-	-	-	3.240	0.914	1.218	99.96			
DTE	MIDAS	0.156	0.360	0.317	0.288	0.197	0.202	-	-	-	-	0.016	-	407.385	-	3.181	0.937	1.264	99.96			
	H-MIDAS	0.213	0.344	0.905	0.309	0.133	0.019	-	-	-	-	0.006	0.041	36.807	880.684	3.062	1.007	1.352	99.97			
	MEM	0.050	0.356	0.803	0.267	0.327	-	-	-	-	-	-	-	-	-	3.152	1.348	1.436	99.98			
	CMEM	0.102	0.356	0.971	0.253	0.289	0.056	0.193	0.745	-	-	-	-	-	-	3.098	1.372	1.538	99.98			
GIA	HAR	0.100	0.356	1.113	0.262	0.282	0.246	-	0.212	0.255	0.251	-	-	-	-	3.227	1.284	1.453	99.98			
	MIDAS	0.170	0.349	1.737	0.252	0.229	0.166	-	-	-	-	0.016	-	336.148	-	3.074	1.397	1.560	99.98			
	H-MIDAS	0.245	0.343	1.839	0.237	0.175	0.092	-	-	-	-	0.004	0.040	82.794	756.965	3.161	1.347	1.495	99.98			
	MEM	0.025	0.303	0.005	0.341	0.331	-	-	-	-	-	-	-	-	-	3.248	0.655	1.076	99.78			
SZG	CMEM	0.075	0.304	-0.010	0.350	0.271	0.005	0.090	0.907	-	-	-	-	-	-	3.374	0.624	1.061	99.79			
	HAR	0.089	0.307	-0.034	0.370	0.235	0.075	-	0.230	0.247	0.444	-	-	-	-	3.373	0.620	1.056	99.79			
	MIDAS	0.155	0.300	-0.035	0.373	0.172	0.096	-	-	-	-	0.017	-	346.460	-	3.401	0.615	1.053	99.79			
	H-MIDAS	0.219	0.294	-0.042	0.355	0.133	0.029	-	-	-	-	0.005	0.041	30.405	602.523	3.407	0.614	1.048	99.79			
VOW	MEM	0.033	0.301	0.008	0.380	0.286	-	-	-	-	-	-	-	-	-	2.520	0.930	1.511	99.85			
	CMEM	0.091	0.292	-0.006	0.382	0.234	0.051	0.195	0.738	-	-	-	-	-	-	2.606	0.885	1.511	99.85			
	HAR	0.086	0.289	0.006	0.377	0.248	0.151	-	0.152	0.410	0.245	-	-	-	-	2.579	0.900	1.533	99.85			
	MIDAS	0.152	0.287	0.015	0.378	0.183	0.151	-	-	-	-	0.016	-	377.033	-	2.615	0.884	1.503	99.85			
YOW	H-MIDAS	0.218	0.279	0.007	0.366	0.137	0.050	-	-	-	-	0.004	0.042	12.822	720.472	2.657	0.866	1.468	99.85			
	MEM	0.023	0.295	0.030	0.360	0.323	-	-	-	-	-	-	-	-	-	2.902	0.618	1.111	99.50			
	CMEM	0.093	0.308	0.023	0.355	0.244	0.018	0.181	0.798	-	-	-	-	-	-	2.972	0.600	1.106	99.49			
	HAR	0.094	0.308	0.019	0.355	0.244	0.084	-	0.182	0.409	0.308	-	-	-	-	2.971	0.600	1.107	99.50			
H-MIDAS	MIDAS	0.167	0.304	0.038	0.338	0.191	0.096	-	-	-	-	0.017	-	295.705	-	2.971	0.598	1.102	99.48			
	H-MIDAS	0.273	0.294	0.031	0.320	0.113	0.032	-	-	-	-	0.006	0.037	36.004	796.936	3.005	0.595	1.094	99.49			

Parameter estimates for the full sample period 02 Jan 2009 - 27 Dec 2012. In **bold** coefficients not significant at 5%. π^\dagger ($\dagger : \times 10^{-2}$)

Table 4: In sample parameter estimates based on the Dynamic ZAF distribution

		Intra-daily parameters					Trend parameters								DZAF parameters							
Ticker	Model	ω^*	α_1	α_0	β_1	β_2	m	$\alpha_{1,d}$	$\beta_{1,d}$	$\beta_{1,w}$	$\beta_{1,m}$	θ_d	θ_h	ω_d	ω_h	a	b	c	ϖ	δ_1	γ_1	
BEI	MEM	0.041	0.302	-0.018	0.353	0.304	-	-	-	-	-	-	-	-	-	-	3.165	0.747	1.192	0.469	0.107	0.938
	CMEM	0.087	0.305	-0.034	0.351	0.257	0.096	0.224	0.669	-	-	-	-	-	-	-	3.167	0.745	1.222	0.477	0.108	0.936
	HAR	0.084	0.305	-0.026	0.347	0.264	0.267	-	0.190	0.350	0.162	-	-	-	-	-	3.181	0.740	1.214	0.488	0.108	0.935
	MIDAS	0.159	0.299	-0.073	0.339	0.203	0.184	-	-	-	-	0.015	-	281.754	-	-	3.189	0.741	1.214	0.464	0.108	0.938
	H-MIDAS	0.235	0.290	-0.109	0.325	0.149	0.124	-	-	-	-	0.006	0.032	129.749	731.739	-	3.208	0.735	1.207	0.479	0.108	0.936
CON	MEM	0.066	0.345	0.262	0.285	0.304	-	-	-	-	-	-	-	-	-	-	3.111	0.945	1.355	0.284	0.092	0.967
	CMEM	0.095	0.362	0.195	0.303	0.241	0.037	0.194	0.779	-	-	-	-	-	-	-	3.216	0.920	1.236	0.356	0.100	0.958
	HAR	0.095	0.360	0.197	0.303	0.241	0.120	-	0.201	0.390	0.327	-	-	-	-	-	3.189	0.929	1.254	0.361	0.097	0.958
	MIDAS	0.158	0.357	0.159	0.292	0.192	0.189	-	-	-	-	0.016	-	418.442	-	-	3.195	0.930	1.256	0.338	0.100	0.960
	H-MIDAS	0.215	0.351	0.302	0.288	0.147	0.026	-	-	-	-	0.006	0.041	36.249	883.925	-	3.153	0.969	1.289	0.365	0.103	0.958
DTE	MEM	0.048	0.356	8.311	0.271	0.324	-	-	-	-	-	-	-	-	-	-	3.096	1.383	1.478	4.218	1.000	0.589
	CMEM	0.099	0.351	2.052	0.265	0.285	0.061	0.188	0.743	-	-	-	-	-	-	-	3.102	1.368	1.538	2.756	0.376	0.725
	HAR	0.102	0.359	2.141	0.259	0.279	0.218	-	0.213	0.385	0.158	-	-	-	-	-	3.048	1.403	1.576	4.584	0.922	0.573
	MIDAS	0.170	0.349	2.455	0.252	0.229	0.168	-	-	-	-	0.016	-	345.092	-	-	3.115	1.370	1.526	4.434	0.989	0.580
	H-MIDAS	0.249	0.342	2.949	0.236	0.173	0.092	-	-	-	-	0.005	0.039	92.903	811.213	-	3.130	1.367	1.518	3.518	0.469	0.667
GIA	MEM	0.025	0.301	-0.001	0.369	0.305	-	-	-	-	-	-	-	-	-	-	3.349	0.628	1.028	0.145	0.118	0.978
	CMEM	0.084	0.304	-0.026	0.379	0.233	0.011	0.153	0.837	-	-	-	-	-	-	-	3.343	0.627	1.069	0.158	0.115	0.976
	HAR	0.088	0.307	-0.038	0.370	0.236	0.075	-	0.231	0.251	0.441	-	-	-	-	-	3.375	0.619	1.056	0.150	0.113	0.977
	MIDAS	0.154	0.301	-0.040	0.364	0.181	0.092	-	-	-	-	0.017	-	334.527	-	-	3.372	0.621	1.067	0.116	0.109	0.982
	H-MIDAS	0.227	0.293	-0.056	0.350	0.131	0.017	-	-	-	-	0.005	0.042	16.616	633.438	-	3.397	0.616	1.054	0.129	0.110	0.980
SZG	MEM	0.033	0.301	-0.007	0.382	0.284	-	-	-	-	-	-	-	-	-	-	2.709	0.847	1.335	0.220	0.133	0.969
	CMEM	0.091	0.292	-0.030	0.381	0.235	0.052	0.199	0.732	-	-	-	-	-	-	-	2.611	0.882	1.507	0.211	0.132	0.970
	HAR	0.087	0.292	-0.029	0.380	0.241	0.174	-	0.148	0.436	0.189	-	-	-	-	-	2.614	0.881	1.504	0.212	0.131	0.970
	MIDAS	0.153	0.287	-0.023	0.376	0.184	0.153	-	-	-	-	0.015	-	398.392	-	-	2.632	0.876	1.489	0.203	0.132	0.971
	H-MIDAS	0.219	0.279	-0.031	0.366	0.136	0.049	-	-	-	-	0.004	0.042	14.204	726.970	-	2.654	0.867	1.471	0.209	0.134	0.970
VOW	MEM	0.023	0.293	0.019	0.360	0.324	-	-	-	-	-	-	-	-	-	-	2.893	0.620	1.119	0.195	0.152	0.965
	CMEM	0.092	0.307	0.004	0.355	0.246	0.018	0.184	0.796	-	-	-	-	-	-	-	2.982	0.597	1.102	0.199	0.149	0.964
	HAR	0.092	0.307	0.001	0.361	0.240	0.080	-	0.181	0.409	0.319	-	-	-	-	-	2.971	0.600	1.106	0.186	0.146	0.967
	MIDAS	0.150	0.304	0.009	0.344	0.202	0.080	-	-	-	-	0.018	-	186.774	-	-	2.988	0.593	1.094	0.201	0.150	0.964
	H-MIDAS	0.271	0.294	0.002	0.324	0.111	0.030	-	-	-	-	0.006	0.037	37.898	795.040	-	3.012	0.592	1.092	0.179	0.145	0.968

Parameter estimates for the full sample period 02 Jan 2009 - 27 Dec 2012. In **bold** coefficients not significant at 5%. For ω^* the average value is shown.

Table 5: Log-likelihood and BIC values for models based on the ZAF and DZAF distributions

		\mathcal{L}_{ZAF}	BIC_{ZAF}	\mathcal{L}_{DZAF}	BIC_{DZAF}	\mathcal{L}^-	BIC^-
BEI	MEM	-20939.5	41963.7	-20907.7	41921.3	31.8	-42.4
	CMEM	-20798.6	41713.6	-20767.4	41672.4	31.2	-41.2
	HAR	-20822.9	41772.8	-20781.6	41711.5	41.2	-61.3
	MIDAS	-20689.9	41496.3	-20656.4	41450.4	33.6	-45.9
	H-MIDAS	-20666.2	41470.0	-20635.0	41428.9	31.2	-41.1
CON	MEM	-18856.7	37795.7	-18811.4	37725.8	45.2	-69.9
	CMEM	-18722.7	37558.5	-18678.1	37489.9	44.6	-68.6
	HAR	-18724.6	37572.7	-18678.8	37501.6	45.9	-71.2
	MIDAS	-18655.5	37424.2	-18613.6	37361.0	41.9	-63.2
	H-MIDAS	-18624.0	37381.2	-18581.0	37315.8	43.0	-65.4
DTE	MEM	-11050.4	22184.9	-10995.3	22095.7	55.1	-89.2
	CMEM	-10916.7	21949.0	-10861.9	21860.4	54.8	-88.6
	HAR	-10936.1	21998.2	-10876.8	21900.6	59.3	-97.6
	MIDAS	-10820.1	21755.8	-10765.0	21666.5	55.2	-89.3
	H-MIDAS	-10771.0	21678.2	-10714.3	21585.8	56.7	-92.4
G1A	MEM	-14257.1	28598.2	-14186.3	28477.7	70.8	-120.5
	CMEM	-14045.5	28206.5	-13973.6	28083.8	71.9	-122.7
	HAR	-14045.3	28216.6	-13979.8	28106.7	65.4	-109.9
	MIDAS	-13971.1	28057.5	-13907.3	27950.8	63.8	-106.7
	H-MIDAS	-13935.6	28007.8	-13866.2	27890.0	69.4	-117.7
SZG	MEM	-11786.7	23656.5	-11709.8	23523.4	76.9	-133.0
	CMEM	-11650.3	23414.9	-11576.7	23288.5	73.6	-126.3
	HAR	-11654.9	23434.4	-11580.5	23306.5	74.3	-127.9
	MIDAS	-11583.1	23280.4	-11509.0	23153.1	74.0	-127.3
	H-MIDAS	-11558.2	23251.5	-11485.1	23126.0	73.2	-125.5
VOW	MEM	-19289.3	38663.4	-19172.2	38450.3	117.1	-213.1
	CMEM	-19005.0	38126.5	-18891.4	37920.5	113.6	-206.0
	HAR	-19010.1	38147.3	-18897.0	37942.3	113.1	-205.0
	MIDAS	-18945.3	38006.9	-18822.3	37782.0	123.0	-224.9
	H-MIDAS	-18876.5	37890.7	-18761.4	37681.5	115.2	-209.2

The table reports the log-likelihoods of models estimated using the ZAF (\mathcal{L}_{ZAF}) and DZAF (\mathcal{L}_{DZAF}) distributions; BICs for ZAF (BIC_{ZAF}) and DZAF (BIC_{DZAF}); the gains in log-likelihood and BIC obtained using the DZAF rather than the ZAF distribution. \mathcal{L}^- : $\mathcal{L}(DZAF) - \mathcal{L}(ZAF)$, positive values are in favour of DZAF; BIC^- : $BIC(DZAF) - BIC(ZAF)$, negative values are in favour of DZAF.

Next, we look at residual diagnostics. For ZAF based models, the autocorrelations of the estimated residuals $\hat{\varepsilon}_{t,i}$ are reported in the left panel of Table 6 with values in boldface indicating a significant Ljung-Box Q-statistic at the same lag. In general, before moving to the discussion of results, it should be noted that, although the residual autocorrelations yielded by the (H)-MIDAS-CMEM and the other fitted models always take values in module very close to zero, the huge number of intra-daily observations makes the test extremely sensitive to deviations from the null hypothesis of white noise errors. More in detail, we find that the only models not exhibiting lack of fit at lag 1 are the MIDAS-CMEM and the H-MIDAS-CMEM. The white noise hypothesis is however almost always rejected at higher lags with two exceptions: the H-MIDAS-CMEM fully captures the residual autocorrelations for SZG, while for CON this happens at lags 1 and 51 (1-day), leaving some noise in the middle of the trading day. The middle panel of Table 6 reports the sample autocorrelations of squared residuals $\hat{\varepsilon}_{t,i}^2$. The analysis essentially confirms the findings for $\hat{\varepsilon}_{t,i}$.

Table 6: Residuals Analysis

Ticker	Model	Residuals $\hat{\varepsilon}_{t,i}$				Residuals $\hat{\varepsilon}_{t,i}^2$				Residuals $\hat{u}_{t,i}$			
		$\hat{\rho}_1$	$\hat{\rho}_{17}$	$\hat{\rho}_{34}$	$\hat{\rho}_{51}$	$\hat{\rho}_1$	$\hat{\rho}_{17}$	$\hat{\rho}_{34}$	$\hat{\rho}_{51}$	$\hat{\rho}_1$	$\hat{\rho}_{17}$	$\hat{\rho}_{34}$	$\hat{\rho}_{51}$
BEI	MEM	0.020	0.000	0.000	0.024	0.021	-0.001	-0.002	0.014	0.001	-0.001	-0.001	-0.001
	CMEM	0.013	0.007	0.001	0.020	0.018	0.002	-0.001	0.014	0.001	-0.001	-0.001	-0.001
	HAR	0.015	0.005	0.000	0.019	0.022	0.002	-0.001	0.015	0.002	-0.001	-0.001	-0.001
	MIDAS	0.004	0.007	-0.004	0.016	0.013	0.002	-0.002	0.012	0.001	-0.001	-0.001	-0.001
	H-MIDAS	-0.002	-0.004	-0.009	0.018	0.008	-0.002	-0.005	0.013	0.001	-0.001	-0.001	-0.001
CON	MEM	0.021	0.002	0.004	0.016	0.007	-0.004	0.013	0.000	0.004	-0.001	-0.001	0.000
	CMEM	0.008	0.005	0.003	0.010	0.004	-0.003	0.016	-0.002	0.002	-0.001	0.000	0.000
	HAR	0.015	0.005	0.004	0.011	0.006	-0.003	0.014	-0.002	0.002	-0.001	0.000	0.000
	MIDAS	-0.002	0.004	-0.002	0.007	0.001	-0.003	0.014	-0.002	0.002	-0.001	-0.001	0.000
	H-MIDAS	-0.002	-0.004	-0.006	0.008	0.001	-0.005	0.013	-0.001	0.000	-0.001	0.000	0.000
DTE	MEM	0.023	0.005	0.012	0.018	0.033	0.003	0.013	0.005	0.000	0.000	0.000	0.000
	CMEM	0.015	0.011	0.012	0.014	0.031	0.007	0.014	0.003	0.000	0.000	0.000	0.000
	HAR	0.016	0.011	0.012	0.013	0.032	0.006	0.013	0.002	0.000	0.000	0.000	0.000
	MIDAS	0.009	0.011	0.006	0.006	0.025	0.008	0.011	-0.001	0.000	0.000	0.000	0.000
	H-MIDAS	0.000	0.000	0.002	0.009	0.017	0.002	0.008	0.001	0.000	0.000	0.000	0.000
G1A	MEM	0.019	-0.007	0.013	0.015	0.005	-0.007	0.004	0.000	0.002	0.004	-0.001	-0.003
	CMEM	0.010	0.001	0.014	0.015	0.003	-0.005	0.004	0.001	0.003	0.004	-0.001	-0.003
	HAR	0.008	0.002	0.014	0.011	0.002	-0.003	0.002	-0.001	0.003	0.004	-0.001	-0.003
	MIDAS	0.000	0.002	0.008	0.004	0.000	-0.004	0.002	-0.002	0.004	0.003	-0.001	-0.003
	H-MIDAS	-0.004	-0.005	0.005	0.009	-0.002	-0.007	0.002	-0.001	0.004	0.004	-0.001	-0.003
SZG	MEM	0.006	-0.003	-0.003	0.017	0.001	0.000	-0.004	0.013	0.002	0.006	-0.003	-0.002
	CMEM	0.004	0.004	0.000	0.013	0.000	0.001	-0.001	0.009	0.003	0.006	-0.003	-0.002
	HAR	0.006	0.003	0.000	0.014	0.002	0.000	-0.001	0.009	0.003	0.006	-0.003	-0.002
	MIDAS	-0.004	0.003	-0.006	0.007	-0.005	0.000	-0.004	0.008	0.003	0.005	-0.003	-0.003
	H-MIDAS	-0.008	-0.005	-0.009	0.009	-0.007	-0.004	-0.004	0.009	0.003	0.006	-0.003	-0.003
VOW	MEM	0.028	0.001	0.003	0.017	0.020	-0.003	-0.004	0.016	0.000	-0.003	0.005	0.013
	CMEM	0.010	0.011	0.005	0.016	0.009	0.001	-0.002	0.018	0.001	-0.003	0.006	0.013
	HAR	0.009	0.011	0.005	0.016	0.009	0.001	-0.003	0.018	0.001	-0.003	0.005	0.013
	MIDAS	0.001	0.010	-0.003	0.006	0.003	0.001	-0.005	0.011	0.001	-0.003	0.006	0.013
	H-MIDAS	-0.007	-0.003	-0.005	0.012	-0.002	-0.003	-0.006	0.016	0.002	-0.003	0.005	0.013

In sample Residuals Analysis for the examined models: full sample period 02 Jan 2009 - 27 Dec 2012. $\hat{\rho}_l$: autocorrelation at the l -th lag; in **bold**: null hypothesis of absence of serial autocorrelation up to the l -th lag is rejected at 5% according to the Ljung-Box test. The Ljung-Box statistics are based on the residuals $\hat{\varepsilon}_{t,i}$ and $\hat{\varepsilon}_{t,i}^2$ for the ZAF-models and on the $\hat{u}_{t,i} = \frac{\mathcal{I}_{t,i} - \hat{\pi}_{t,i}}{\sqrt{\hat{\pi}_{t,i}(1 - \hat{\pi}_{t,i})}}$ of the ACM component for the DZAF-models.

Finally, since for the DZAF specifications the $\varepsilon_{t,i}$ are independently but not identically distributed, for these models the diagnostics rely on the sample autocorrelations of the residuals of the ACM component,

$$\hat{u}_{t,i} = (\mathcal{I}_{t,i} - \hat{\pi}_{t,i}) / \sqrt{\hat{\pi}_{t,i}(1 - \hat{\pi}_{t,i})},$$

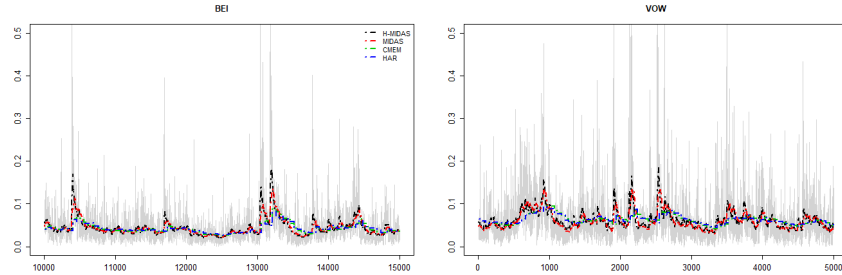
whose values are reported in the last panel of Table 6. In this case, the null hypothesis of white noise errors can be never rejected except for VOW at lag 51. This means that the ACM(1,1) specification fully captures the serial dependence in the trade indicator dynamics. Finally, as the autocorrelations for $\hat{u}_{t,i}^2$ are negligible, to save space, their values are not reported in the paper.

5.3. Fitted dynamic components

In this section we provide further insight on the statistical properties of the dynamic components fitted using the H-MIDAS-CMEM and the competing models considered. We focus on models with ZAF errors, omitting to report and comment the findings for DZAF models, since they do not substantially differ from those obtained for models with ZAF errors. The corresponding tables and plots for DZAF models are however made available in the online Empirical Appendix.

First of all, it is interesting to focus on the graphical analysis of the estimated trends. In particular, this is done in Figure 3 where we report the time plots of the fitted long-run components obtained from models with ZAF errors. For reasons of space and clarity, we confine our attention to two different subsamples of 5000 observations for the stocks BEI and VOW². The plots reveal that, for the MIDAS-CMEM and, in particular, for the H-MIDAS-CMEM specifications, the fitted trend components are more reactive to shocks in the intra-daily volumes than their competitors CMEM and HAR-CMEM. This is evidently due to the fact that, for these two models, the trend component varies on a daily scale, being fixed within a given trading day.

Figure 3: Models with ZAF errors. Fitted long-run components vs seasonally adjusted volumes over a subsample of 5000 observations for BEI and VOW.



Key to figure: CMEM (green), HAR-CMEM (blue), MIDAS-CMEM (red) and H-MIDAS-CMEM (black). Seasonally adjusted intra-daily volumes are drawn in grey (for each series values are normalised with respect to the maximum observed trading volume).

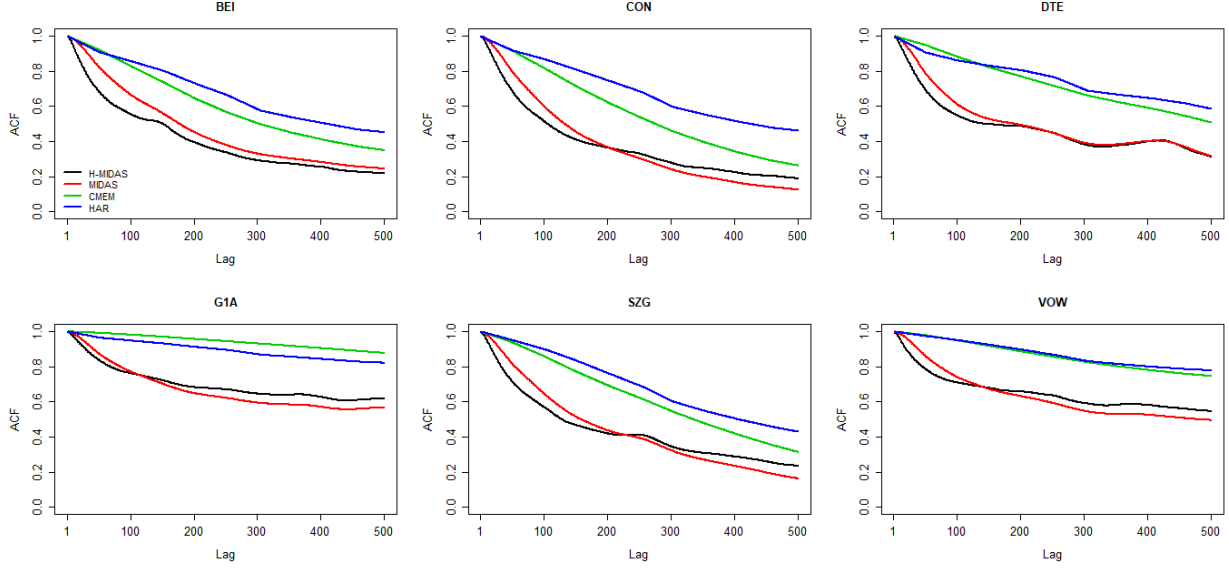
Next, in Figure 4, we provide some insight on the effect that, in the H-MIDAS-CMEM, the introduction of the hourly MIDAS filter has on the dynamic properties of the fitted long-run component. Here, the analysis of the autocorrelation functions of the estimated long-run components up to 500 lags, approximately corresponding to two trading weeks, shows that the H-MIDAS trend is less strongly autocorrelated than the MIDAS one for lower lags. However, at higher lags, the ACFs of the $\tau_{t,i}$ components fitted by H-MIDAS filters, in four cases out of six, tends to decay more slowly than the ACFs of the long-run components fitted by the *plain* daily MIDAS filters. For completeness, we also report the autocorrelation patterns of the long-run components implied by the HAR-CMEM and CMEM, finding that, as expected, these are remarkably more persistent than what found for the MIDAS based specifications.

In Figure 5, focusing on the H-MIDAS, we compare the sample autocorrelation functions of the fitted $\tau_{t,i}$ and $g_{t,i}$. Our aim is here to highlight the distinct and separate contributions that the short and long-run components provide to the modelling of the dynamics of the seasonally adjusted intra-daily volumes. Under this respect, the plots make evident that the two components are characterised by remarkably different autocorrelation patterns with $g_{t,i}$ only accounting for short term movements around the long-run component and $\tau_{t,i}$ capturing longer term movements.

Finally, we analyse the weighting functions characterising the MIDAS filters involved in the long-run components of the fitted H-MIDAS-CMEMs. As suggested by the estimation results, the shapes of these functions are remarkably different across assets. This is clearly visible in figures 6 and 7, reporting the shapes of the estimated weighting functions for the daily and hourly components, respectively. In order to summarise the information contained in the plots, Table 7 reports the number of daily (2 hour 50 minute) lags needed for the weights of the daily (hourly) filter to

²The time plots of the fitted long-run components for all the stocks and the full sample period are available in Figure 10 in the online Empirical Appendix.

Figure 4: Models with ZAF errors. Sample ACFs of the long-run components up to lag 500.



Key to figure: CMEM (green), HAR-CMEM (blue), MIDAS-CMEM (red) and H-MIDAS-CMEM (black). Models have been estimated by the ZAF distribution.

reach 10^{-2} and 10^{-6} , respectively. For the daily component, the fastest decay takes place for the BEI stock, for which the weights decline to 10^{-2} in 7 trading days, while they take 23 trading days to reach 10^{-6} . The slowest decay is observed for SZG, for which the decay times needed to reach the same values are equal to 37 and 222 trading days, respectively. As expected, the weights of the hourly component decline to zero more rapidly with the fastest decay observed for VOW, taking 4 periods (approximately 11 trading hours) to reach 10^{-2} and 13 periods (approximately 37 trading hours), to reach 10^{-6} . The slowest decays occur for SZG and CON, requiring 7 periods (approximately 20 trading hours) to reach 10^{-2} and 21 periods (approximately 60 trading hours) to reach 10^{-6} .

5.4. Out-of-sample forecasting

To evaluate the predictive ability of the H-MIDAS-CMEM model and its relative merits with respect to the competitors, we have performed a forecasting comparison over the out-of-sample period 17 December 2010 – 27 December 2012. In order to capture the salient features of the data and safeguard against the presence of structural breaks, the model parameters have been recursively estimated every day over a 500-day rolling window. Therefore, at each

Table 7: Fitted H-MIDAS-CMEM models with ZAF errors. Number of days taken by the Beta weights of the “daily” filter ($\varphi_k(\omega_d)$) and number of 2 hour 50 minute-intervals taken by the Beta weights of the “hourly” filter ($\varphi_{l,k}(\omega_h)$) to decay to 10^{-2} and 10^{-6} , respectively.

	Decay time to 10^{-2}		Decay time to 10^{-6}	
	daily	hourly	daily	hourly
BEI	7	5	23	14
CON	26	7	124	21
DTE	12	6	45	17
G1A	23	6	97	20
SZG	37	7	222	21
VOW	18	4	69	13

Figure 5: H-MIDAS-CMEM models with ZAF errors. Sample autocorrelation functions of the components $g_{t,i}$ (in red) and $\tau_{t,i}$ (in black) (up to 1 day).

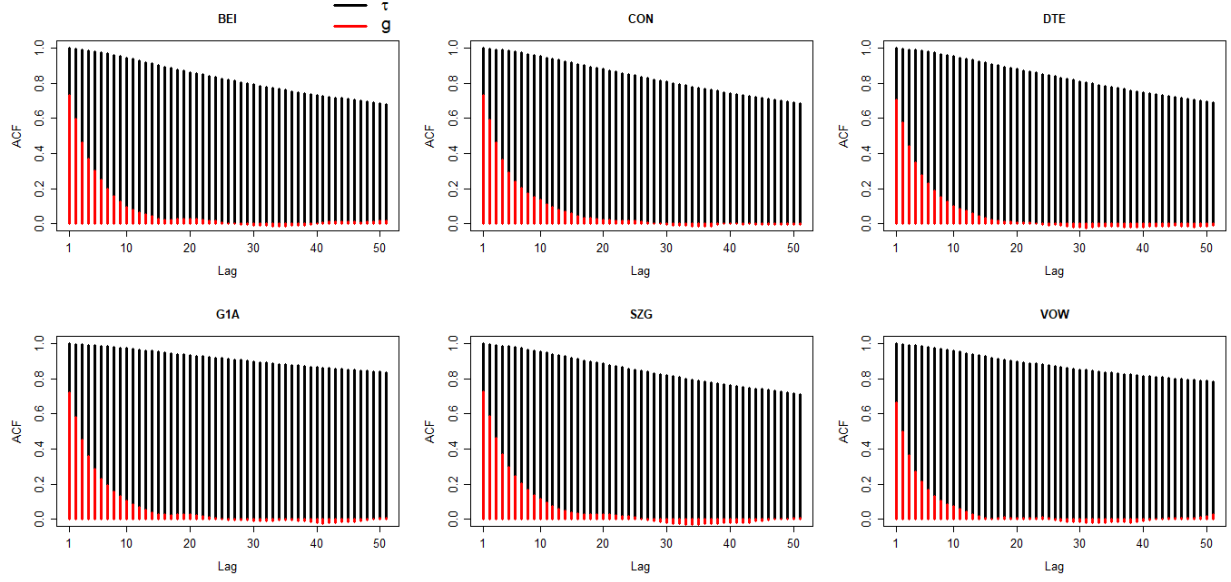
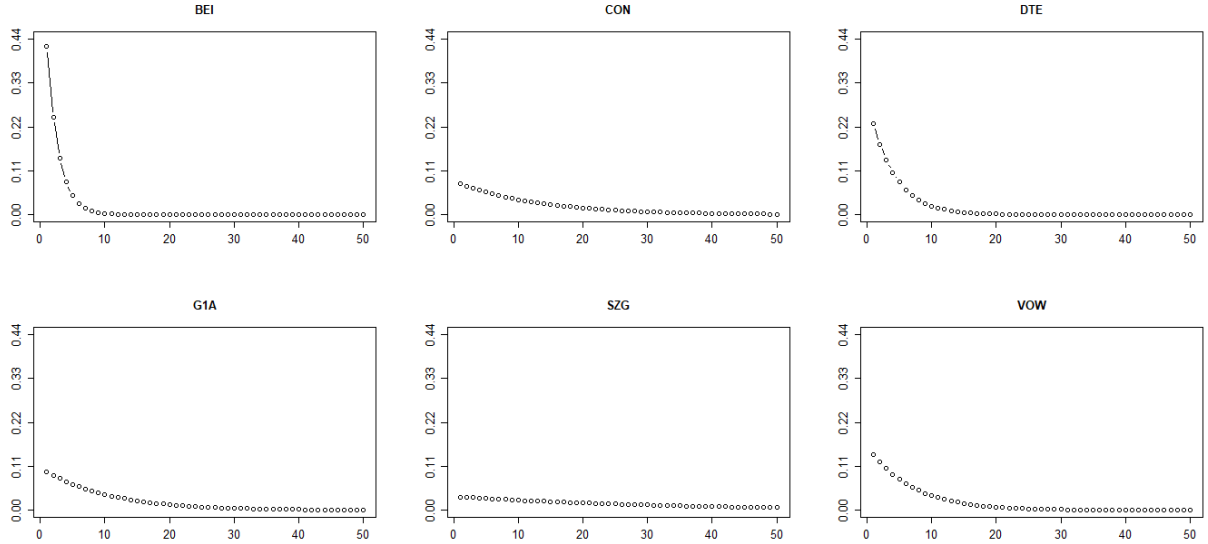
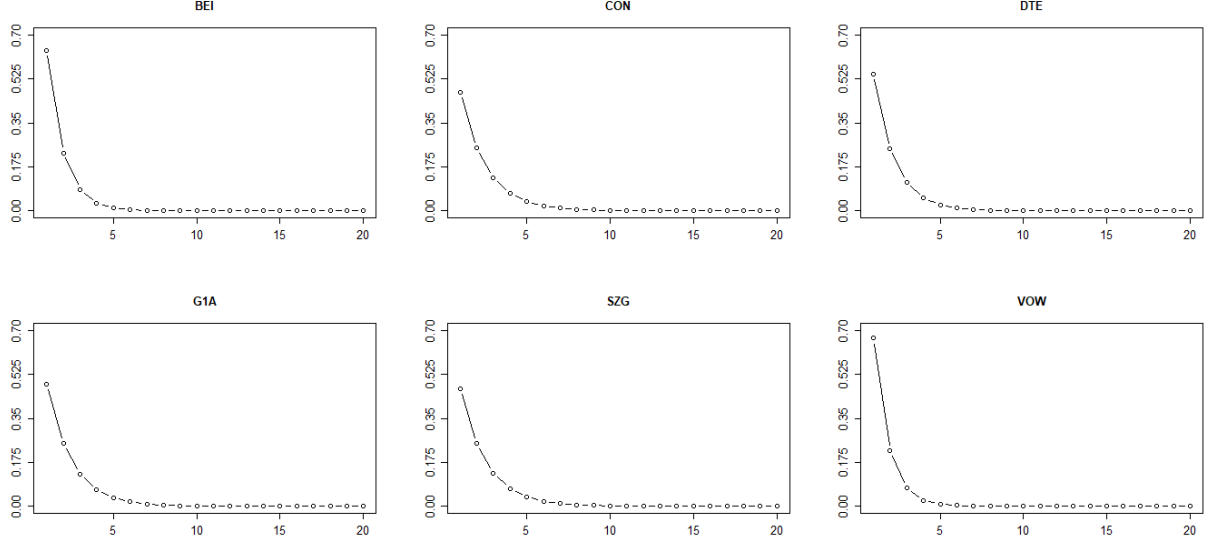


Figure 6: Decay profile of the Beta weighting function of the daily MIDAS filter for the H-MIDAS-CMEM with ZAF errors: beta weights (vertical axis) vs. daily lags (horizontal axis).



step we have predicted 51 intra-daily volumes before re-estimating all the models, for a total of 517 days and 26367 intra-daily observations included in our out-of-sample period. The predictive performance of the examined models has been evaluated by computing some widely used forecasting loss functions. The significance of differences in forecasting performance has been assessed by means of the Model Confidence Set (MCS) approach ([Hansen et al.](#),

Figure 7: Decay profile of the Beta weighting function of the hourly MIDAS filter for the H-MIDAS-CMEM with ZAF errors: beta weights (vertical axis) vs. hourly lags (horizontal axis).



2011), which relies on a sequence of statistical tests in order to identify, at a certain confidence level $(1 - \alpha)$, the set of superior models with respect to some appropriately chosen measure of predictive ability.

To compare the out-of-sample predictive performance of the models, we have considered the following loss functions

$$MSE = \sum_{t=1}^T \sum_{i=1}^I (x_{t,i} - \hat{x}_{t,i})^2, \quad (37)$$

$$MAE = \sum_{t=1}^T \sum_{i=1}^I |x_{t,i} - \hat{x}_{t,i}|, \quad (38)$$

$$Slicing = - \sum_{t=1}^T \sum_{i=1}^I w_{t,i} \log \hat{w}_{t,i}, \quad (39)$$

where MSE is the Mean Squared Error, MAE is the Mean Absolute Error and $Slicing$ is the Slicing loss function. This has been developed by Brownlees et al. (2011) for evaluating trading strategies based on the replication of the Volume Weighted Average Price (VWAP). The use of this loss function can be motivated from an operational as well as a statistical point of view.

From an operational point of view, it is worth reminding that a key objective in high frequency trading is to minimise transaction costs of a given order by optimally slicing it during the day, aiming to achieve an average execution price as close as possible to the VWAP. Introduced by Berkowitz et al. (1988) as an unbiased estimate of prices which could be reached by any non-strategic trader, the VWAP is defined as the ratio between the total traded value and the total traded volume (Madhavan, 2002). Formally, the VWAP for day t can be expressed as

$$VWAP_t = \frac{\sum_{j=1}^{N_t} v_t^{(j)} p_t^{(j)}}{\sum_{j=1}^{N_t} v_t^{(j)}}, \quad (40)$$

where N_t denotes the number of transactions within the day t , with $p_t^{(j)}$ and $v_t^{(j)}$ being the price and the volume of the j -th transaction of the t -th day, respectively. Considering I equally spaced intervals during the trading day, Eq. (40)

can be rewritten as

$$VWAP_t = \frac{\sum_{i=1}^I \left(\sum_{j \in B_i} v_t^{(j)} \right) \bar{p}_{t,i}}{\sum_{i=1}^I \left(\sum_{j \in B_i} v_t^{(j)} \right)} = \frac{\sum_{i=1}^I x_{t,i} \bar{p}_{t,i}}{\sum_{i=1}^I x_{t,i}} = \sum_{i=1}^I w_{t,i} \bar{p}_{t,i}, \quad (41)$$

with $\bar{p}_{t,i}$ being the VWAP of the i -th partition of the trading day, $w_{t,i}$ the corresponding intra-daily proportion of traded volumes and B_i indicating the set of transactions falling within the i -th partition. Namely,

$$w_{t,i} = \frac{x_{t,i}}{\sum_{i=1}^I x_{t,i}}. \quad (42)$$

Similarly, let us define the Average Execution Price (AEP) over day t as

$$AEP_t = \sum_{i=1}^I w_{t,i}^* \bar{p}_{t,i},$$

where $w_{t,i}^*$ ($i = 1, \dots, I$) indicates an arbitrarily defined order slicing strategy over day t . It can be shown by simple algebra that the discrepancy between AEP_t and $VWAP_t$ is minimised by choosing $w_{t,i}^* = w_{t,i}$, suggesting that accurate prediction of volume proportions is of fundamental importance for VWAP traders.

The next step is to provide a statistical justification for the use of the *Slicing* defined in (39) as a tool for assessing the accuracy of forecasts of intra-daily volume proportions as well as for indirectly assessing the accuracy of VWAP replication strategies. Under this respect, it is important to note that the Slicing loss in Eq. (39), as discussed in Brownlees et al. (2011), is related to the Kullback-Leibler discrepancy between the distributions of observed and forecast intra-daily volumes. In addition, it can be shown to be equal to the leading term of a negative predictive multinomial log-likelihood. Again, details and analytical derivations are provided in Brownlees et al. (2011).

Depending on how the slicing weights $\hat{w}_{t,i}$ are computed, the Slicing loss can be used to implement two different VWAP replication strategies that will be denoted as *static* and *dynamic*, respectively. In the static case the i -th weight represents the i -th intra-daily volume proportion of day t

$$\hat{w}_{t,i|t-1} = \frac{\hat{x}_{t,i|t-1}}{\sum_{i=1}^I \hat{x}_{t,i|t-1}} \quad i = 1, \dots, I,$$

where $\hat{x}_{t,i|t-1}$ is the forecast of the trading volume over sub-interval i of day t conditional on previous day information. Differently, in the dynamic VWAP replication strategy, the slicing weights are dynamically updated through the formula

$$\hat{w}_{t,i|i-1} = \begin{cases} \frac{\hat{x}_{t,i|i-1}}{\sum_{j=1}^I \hat{x}_{t,j|i-1}} \left(1 - \sum_{j=1}^{i-1} \hat{w}_{t,j|i-1} \right) & i = 1, \dots, I-1, \\ \left(1 - \sum_{i=1}^{I-1} \hat{w}_{t,i|i-1} \right) & i = I. \end{cases} \quad (43)$$

It is worth noting that the Slicing loss provides a criterion for evaluating the effectiveness of trading strategies in reaching the VWAP target through the evaluation of the accuracy of predicted volume proportions. On the other hand, MSE and MAE focus on the evaluation of the accuracy of predicted levels of trading volumes.

Table 8 reports the results of the forecasting evaluation for a horizon equal to 10 minutes (one-step-ahead). In order to facilitate the interpretation of results, we take the MEM with ZAF errors as a benchmark and, for any model M and loss function F , report the relative gain over the benchmark measured in terms of the ratio $100(L_{F,M}/L_{F,MEM-ZAF})$, where $L_{F,M}$ is the value of loss F recorded for model M , so that values lower than 100 will indicate that model M is outperforming the benchmark. Values in boldface indicate the models returning the minimum average loss, while values shaded in grey are associated to models included in the 75% MCS. In the MCS implementation we have considered a Semi-Quadratic statistic³ and 5000 Bootstrap resamples generated by means of a block-bootstrap procedure. The optimal block length has been estimated through the method described in Patton et al. (2009). The full set of MCS p-values is reported in the online Empirical Appendix in Table 14.

³We report p-values only for Semi-Quadratic statistic because the results corresponding to the Range statistic are very similar.

Table 8: Out-of-sample loss functions comparison for ZAF and DZAF

	Relative gain vs MEM-ZAF model											
	BEI				CON				DTE			
	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>
MEM-ZAF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MEM-DZAF	100.08	100.01	100.00	100.00	96.26	98.65	99.98	99.98	99.71	99.34	100.02	100.01
CMEM-ZAF	98.87	97.93	99.93	99.93	97.52	97.69	99.92	99.93	98.78	94.08	99.90	99.90
CMEM-DZAF	98.96	97.94	99.93	99.93	91.43	97.32	99.93	99.94	98.80	97.03	99.96	99.96
HAR-ZAF	99.20	98.02	99.94	99.94	98.35	97.99	99.95	99.95	96.78	93.65	99.89	99.90
HAR-DZAF	99.24	98.02	99.94	99.94	94.94	97.58	99.95	99.96	97.77	96.69	99.96	99.96
MIDAS-ZAF	94.50	96.60	99.89	99.89	92.36	100.69	100.00	100.00	92.01	97.44	99.93	99.94
MIDAS-DZAF	94.74	96.71	99.89	99.89	86.79	96.21	99.90	99.91	91.28	94.98	99.93	99.93
H-MIDAS-ZAF	91.30	95.49	99.85	99.85	82.68	95.49	99.90	99.90	85.18	93.51	99.90	99.90
H-MIDAS-DZAF	91.60	95.59	99.86	99.86	82.57	95.06	99.88	99.89	86.38	93.24	99.90	99.90

	GIA				SZG				VOW			
	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>
	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>
MEM-ZAF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MEM-DZAF	100.20	100.02	100.00	100.00	99.96	99.95	100.02	100.02	100.18	100.00	100.00	100.00
CMEM-ZAF	98.43	97.26	99.91	99.91	97.48	97.51	99.92	99.93	98.36	95.66	99.85	99.86
CMEM-DZAF	98.41	97.28	99.91	99.92	97.84	97.86	99.94	99.94	98.32	95.63	99.85	99.86
HAR-ZAF	98.15	97.11	99.91	99.91	97.84	97.87	99.94	99.94	98.48	95.81	99.86	99.87
HAR-DZAF	98.12	97.12	99.91	99.91	98.09	98.10	99.95	99.95	98.45	95.79	99.86	99.86
MIDAS-ZAF	93.19	95.50	99.84	99.84	96.25	96.28	99.88	99.88	93.88	94.63	99.80	99.81
MIDAS-DZAF	92.99	95.46	99.84	99.84	96.58	96.61	99.88	99.89	93.92	94.59	99.79	99.81
H-MIDAS-ZAF	89.79	94.34	99.79	99.80	94.99	95.01	99.82	99.83	89.83	92.73	99.71	99.72
H-MIDAS-DZAF	89.82	94.36	99.79	99.80	95.64	95.67	99.84	99.85	89.75	92.71	99.70	99.72

The table shows the ratio of the loss functions between all analysed models and MEM-ZAF (benchmark model) considering the ZAF and DZAF distribution. Values smaller than 100 denote improvements over the benchmark. In **bold** the best model and in **box** model \in 75% MCS. Loss functions: Mean Squared Error (*MSE*), Mean Absolute Error (*MAE*) and *Slicing* loss with weights computed under the static (*SL_{stc}*) and dynamic (*SL_{dyn}*) VWAP replication strategy.

It can be easily seen that the H-MIDAS-CMEM specifications are the only ones entering the 75% MCS in almost all cases. The only exception is represented by the DTE stock, for which we find that, for MAE and the Slicing loss functions, the CMEM and HAR-CMEM with ZAF errors enter the MCS together with the two H-MIDAS-CMEM models. Overall, the models with ZAF and DZAF errors return very close performances. So the choice of the error distribution does not appear to be critical for the data at hand.

Furthermore, we have performed an additional forecasting experiment aimed at assessing the ability of the fitted models to forecast trading volumes at longer lead times. Every day, at the market opening, we compute predictions of the trading volume at different horizons, corresponding to different periods of the day. The accuracy of forecasts is assessed by means of MSE and MAE and the significance of loss differentials is tested by the MCS procedure. Namely, we partition the trading day as follows: 9:00 + 9:30, 9:30 + 10:00, 10:00 + 11:00, 11:00 + 13:00, 13:00 + 17:30. The results are reported in Table 9, for the MSE, and in Table 10, for the MAE. As already found for 1-step-ahead forecasts, models with ZAF and DZAF errors return very close performances. The H-MIDAS-CMEM models always enter the MCS except for the MAE loss function computed over the first three sub-intervals for the DTE stock. On the other hand, their competitors are included in the MCS in a few isolated cases: CON and DTE, for the MAE, and SZG, for the MSE. As for the 1-step-ahead case, the associated MCS p-values are reported in the online Empirical Appendix in tables 15 and 16.

The results of the out-of-sample forecasting comparison mostly confirm the findings of the full-sample analysis. First, the MIDAS-CMEM and H-MIDAS-CMEM, updating the trend component intra-daily, tend to outperform their competitors, updating the trend component daily (CMEM, HAR) or being characterised by a constant trend (MEM). Second, in most cases, the H-MIDAS-CMEM is performing significantly better than the MIDAS-CMEM, thus supporting the inclusion of a hourly component in the model for $\tau_{t,i}$. In other words, considering that the MIDAS-CMEM

Table 9: Out-of-sample MSE comparison for different time horizons for ZAF and DZAF

Relative gain vs MEM-ZAF for MSE															
	BEI					CON					DTE				
	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30
MEM-ZAF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MEM-DZAF	99.82	99.82	99.94	99.97	100.06	96.65	91.46	95.32	95.72	95.94	98.43	94.66	94.01	97.62	97.00
CMEM-ZAF	97.92	98.98	100.48	98.34	99.85	98.46	97.86	99.60	99.75	99.25	96.32	96.57	100.37	103.27	99.94
CMEM-DZAF	97.96	98.97	100.48	98.41	99.87	94.97	90.64	91.78	90.69	93.14	98.91	96.77	98.74	99.89	97.86
HAR-ZAF	98.66	99.39	100.57	98.86	99.96	101.86	97.48	99.34	99.62	96.14	98.81	96.05	96.45	98.66	96.76
HAR-DZAF	98.69	99.37	100.62	98.87	99.95	95.74	93.91	93.42	91.25	92.63	98.45	95.21	96.82	98.93	96.02
MIDAS-ZAF	95.24	96.10	99.47	98.35	97.05	94.17	87.12	92.77	95.47	105.68	97.83	93.79	96.84	98.73	103.45
MIDAS-DZAF	95.19	96.00	99.40	98.44	97.44	89.93	87.18	89.47	89.09	91.24	96.41	94.09	96.29	96.09	90.22
H-MIDAS-ZAF	92.25	93.59	97.62	96.24	95.85	82.20	80.89	84.88	86.55	88.41	91.83	87.87	88.56	90.38	82.73
H-MIDAS-DZAF	92.26	93.41	97.72	96.53	95.87	83.26	81.85	86.69	87.00	89.49	93.84	92.05	94.89	94.75	83.31

	G1A					SZG					VOW				
	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30
MEM-ZAF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MEM-DZAF	100.12	100.12	100.04	100.06	99.99	100.16	99.39	99.21	99.55	98.14	100.09	100.04	100.03	99.95	99.89
CMEM-ZAF	97.06	94.37	96.38	99.45	96.71	99.43	98.05	102.11	101.10	97.57	96.09	95.53	96.70	97.77	92.28
CMEM-DZAF	97.05	94.35	96.37	99.41	96.71	99.34	97.03	100.34	100.17	96.94	96.28	95.58	96.66	97.65	92.17
HAR-ZAF	97.35	94.86	96.30	99.21	96.47	99.09	95.24	100.43	100.47	97.75	96.40	96.11	96.99	97.82	92.35
HAR-DZAF	97.35	94.83	96.28	99.16	96.48	98.70	96.80	100.78	100.57	97.70	96.59	96.17	96.96	97.71	92.24
MIDAS-ZAF	93.24	93.11	94.89	98.66	93.01	94.95	95.94	100.22	100.22	96.08	92.46	94.25	97.00	96.62	90.45
MIDAS-DZAF	93.16	92.81	94.77	98.56	93.23	95.01	94.95	100.45	100.58	95.99	92.58	94.11	97.07	96.58	90.41
H-MIDAS-ZAF	91.27	92.06	92.73	95.81	92.33	93.75	95.26	99.19	98.43	93.96	90.31	92.42	95.22	94.68	88.34
H-MIDAS-DZAF	91.26	92.02	92.72	95.84	92.30	93.82	94.40	99.74	99.70	94.46	90.37	92.46	95.25	94.62	88.30

The table shows the ratio of the MSE loss between analysed models and MEM-ZAF (benchmark model). Values smaller than 100 denote improvements over the benchmark. In **bold** the best model; in **box** model \in 75% MCS. The results are based on the use of ZAF and DZAF distribution.

Time horizons: 9:30 = 9:00 + 9:30; 10:00 = 9:30 + 10:00; 11:00 = 10:00 + 11:00; 13:00 = 11:00 + 13:00; 17:30 = 13:00 + 17:30.

Table 10: Out-of-sample MAE comparison for different time horizons for ZAF and DZAF

Relative gain vs MEM-ZAF for MSE															
	BEI					CON					DTE				
	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30
MEM-ZAF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MEM-DZAF	99.97	99.93	99.93	99.97	100.03	99.07	97.64	97.54	98.18	97.52	98.74	96.75	95.97	97.76	97.66
CMEM-ZAF	97.08	96.21	95.60	94.81	96.25	98.06	97.75	98.00	98.22	98.73	91.23	89.15	90.01	92.33	90.86
CMEM-DZAF	97.12	96.27	95.56	94.86	96.23	99.01	98.24	98.46	98.63	99.20	96.99	94.27	93.91	95.19	93.85
HAR-ZAF	97.42	96.29	95.58	95.11	96.22	99.14	97.89	97.97	98.54	97.38	94.13	90.19	90.25	91.85	90.51
HAR-DZAF	97.49	96.31	95.59	95.13	96.23	99.07	98.72	98.72	98.55	98.35	97.70	94.03	93.43	94.77	93.21
MIDAS-ZAF	94.58	94.58	94.55	94.70	94.88	100.17	102.05	105.48	107.18	112.11	101.80	98.93	100.06	101.20	102.42
MIDAS-DZAF	94.53	94.62	94.42	94.95	95.19	97.13	97.77	98.49	98.22	97.56	96.25	93.84	93.90	94.45	91.98
H-MIDAS-ZAF	93.17	93.81	93.88	93.51	93.88	95.95	96.90	98.00	97.84	96.82	95.83	92.81	93.10	93.80	91.59
H-MIDAS-DZAF	93.25	93.83	93.88	93.66	93.85	95.71	96.74	97.67	97.68	96.84	94.52	92.52	92.78	92.94	89.00

	G1A					SZG					VOW				
	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30
MEM-ZAF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MEM-DZAF	99.99	99.99	99.94	99.93	99.90	100.10	99.86	99.68	99.61	99.46	99.96	99.91	99.90	99.86	99.84
CMEM-ZAF	96.09	94.09	93.55	93.72	92.27	96.81	95.02	94.51	94.17	92.79	94.93	92.37	91.49	90.26	87.04
CMEM-DZAF	96.12	94.14	93.60	93.75	92.33	96.95	95.23	94.71	94.43	92.92	95.01	92.42	91.51	90.25	87.03
HAR-ZAF	96.15	94.27	93.54	93.49	92.20	97.00	94.93	94.51	94.32	92.99	95.30	92.59	91.77	90.45	87.23
HAR-DZAF	96.19	94.33	93.58	93.52	92.26	96.92	95.35	95.00	94.74	93.30	95.38	92.64	91.79	90.43	87.22
MIDAS-ZAF	94.42	92.90	92.63	93.27	90.58	94.13	93.51	93.60	93.65	91.98	93.51	92.02	91.70	90.30	86.68
MIDAS-DZAF	94.52	92.78	92.67	93.12	90.58	94.41	93.66	93.93	94.09	92.17	93.57	91.84	91.75	90.27	86.63
H-MIDAS-ZAF	93.09	91.49	91.40	91.71	89.67	93.39	92.84	93.30	92.90	90.90	92.02	90.21	90.09	88.37	84.32
H-MIDAS-DZAF	93.12	91.56	91.47	91.77	89.66	93.64	93.31	93.72	93.54	91.08	92.07	90.26	90.13	88.38	84.33

The table shows the ratio of the MAE loss function between analysed models and MEM-ZAF (benchmark model). Values smaller than 100 denote improvements over the benchmark. In **bold** the best model; in **box** model \in 75% MCS. The results are based on the use of ZAF and DZAF distribution. Time horizons: 9:30 = 9:00 + 9:30; 10:00 = 9:30 + 10:00; 11:00 = 10:00 + 11:00; 13:00 = 11:00 + 13:00; 17:30 = 13:00 + 17:30.

is a restricted version of the H-MIDAS-CMEM, the results of the out-of-sample forecasting comparison indirectly provide evidence against the restriction $\theta_h = 0$.

These findings lead us to the intuition that the successful performance of the proposed H-MIDAS-CMEM is mainly driven by two key factors. The first one is related to the fact that the trend component is updated at intra-daily frequencies. The worst performing models are indeed characterised by a slow-moving trend component that is updated daily or even kept constant. The second one can be identified in the more flexible weighting structure of the H-MIDAS-CMEM compared to the single filter MIDAS-CMEM.

It is also interesting to see that the three loss functions used for our forecasting comparison are not equally sensitive to the choice of the forecasting model. In particular, for MAE and MSE, the performance gaps, between the H-MIDAS-CMEM and its competitors, are remarkable at any forecasting horizon, in some cases being close to 20% over the benchmark. Differently, focusing on 10-minute-ahead predictions, we find that, in terms of the Slicing loss, the observed performance gaps, although still being statistically significant, are much less pronounced. This suggests that, in practical applications, the choice of the forecasting model is particularly important when one is interested in predicting future levels of trading volumes, as assessed by the MSE and MAE. On the other hand, this choice becomes less critical when the main aim is to generate forecasts of volume proportions, whose accuracy can be assessed through the Slicing loss function.

6. Conclusions

The paper introduces the Heterogeneous MIDAS Component MEM (H-MIDAS-CMEM) as a novel approach for fitting and forecasting high frequency volumes. The structure of the model is motivated by the main stylised facts, arising from the empirical analysis of time series of high frequency volumes. Namely, extending the logic of the CMEM developed in [Brownlees et al. \(2011\)](#) by the use of an Heterogeneous-MIDAS component, specified as an additive cascade of linear filters moving at different frequencies, we are able to better capture the main empirical properties of intra-daily trading volumes, such as memory persistence and clustering of the trading activity. In addition, we investigate the statistical properties of the proposed model deriving conditions for stationarity and ergodicity. Inference is performed by means of a two-stage approach.

On the empirical ground, from the analysis of the six German stocks considered, it arises that the H-MIDAS-CMEM provides a very good fit in the in-sample estimation, using both the ZAF and DZAF distributions ([Hautsch et al., 2014](#)). The out-of-sample analysis confirms the strength of the H-MIDAS-CMEM, since it significantly outperforms the competitors in terms of three different loss functions.

A natural extension of the research carried out in this paper would be to construct multivariate specifications that can jointly model the co-movements of trading volumes for a panel of stocks. In addition, another potential enhancement would be to extend the proposed heterogeneous MIDAS framework to account for the impact that suitably chosen financial or economic variables can have on the long-run component.

Online Empirical Appendix

Figure 8: Intra-daily trading volumes for the sample period 02 January 2009 – 27 December 2012 (for each series values are normalised with respect to the maximum trading volume observed over the sample period).

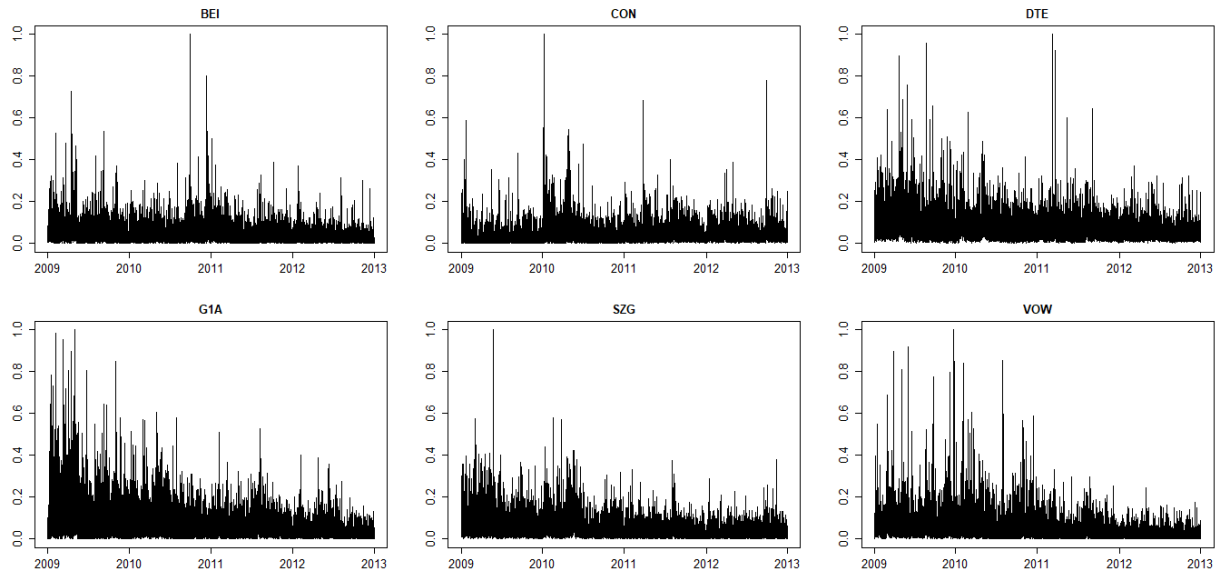
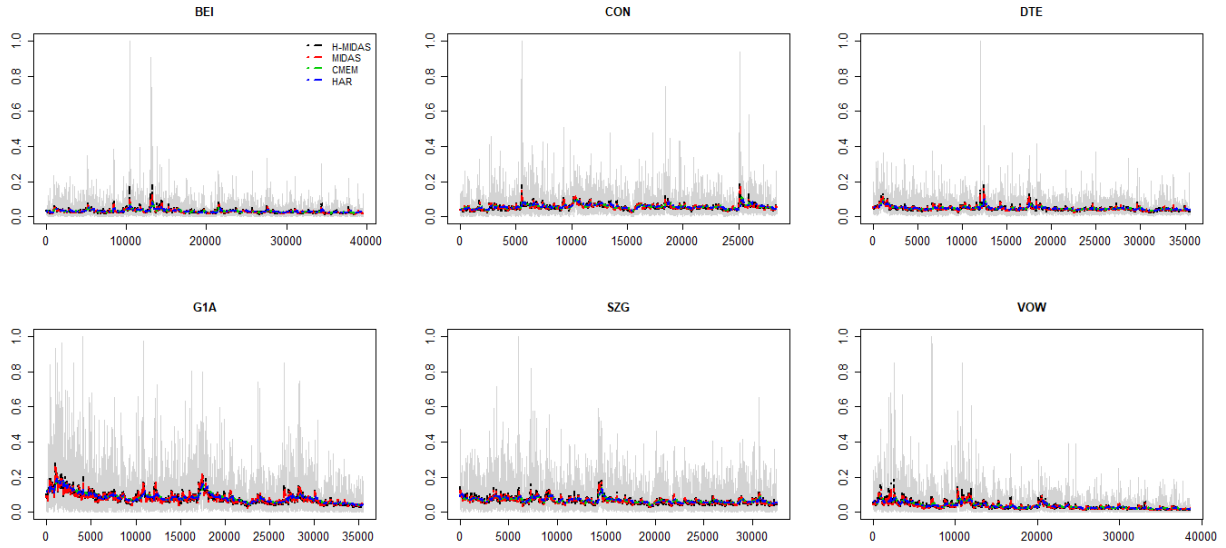
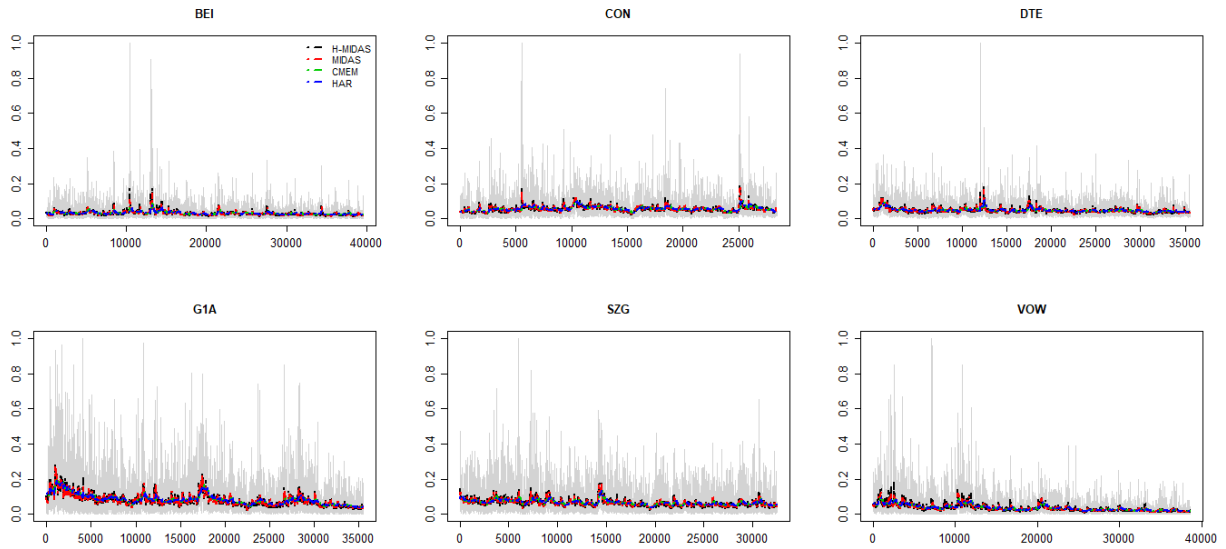


Figure 9: Models with ZAF errors. Fitted long-run components vs seasonally adjusted volumes.



Key to figure: CMEM (green), HAR-CMEM (blue), MIDAS-CMEM (red) and H-MIDAS-CMEM (black). Seasonally adjusted intra-daily volumes are drawn in grey (for each series values are normalised with respect to the maximum observed trading volume).

Figure 10: Models with DZAF errors. Fitted long-run components vs seasonally adjusted volumes.



Key to figure: CMEM (green), HAR-CMEM (blue), MIDAS-CMEM (red) and H-MIDAS-CMEM (black). Seasonally adjusted intra-daily volumes are drawn in grey (for each series values are normalised with respect to the maximum observed trading volume).

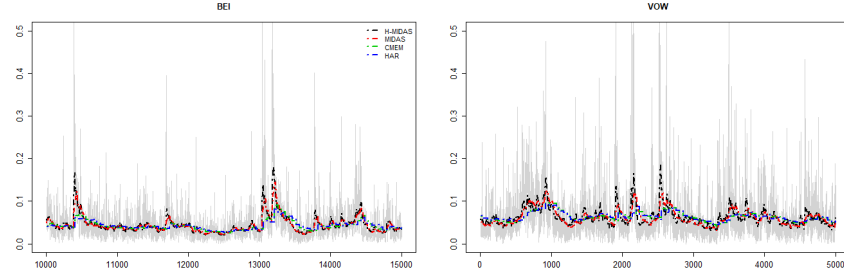
Table 11: In sample parameter estimates using the ZAF distribution: standard errors.

Ticker	Model	Intra-daily parameters				Trend parameters										ZAF parameters			
		α_1	α_0	β_1	β_2	m	$\alpha_{1,d}$	$\beta_{1,d}$	$\beta_{1,w}$	$\beta_{1,m}$	θ_d	θ_h	ω_d	ω_h	a	b	c	π	
BEI	MEM	0.007	0.043	0.016	0.016	-	-	-	-	-	-	-	-	-	0.172	0.052	0.099	0.000	
	CMEM	0.007	0.054	0.017	0.017	0.016	0.022	0.034	-	-	-	-	-	-	0.103	0.032	0.062	0.000	
	HAR	0.007	0.088	0.016	0.017	0.038	-	0.033	0.045	0.055	-	-	-	-	0.175	0.045	0.086	0.000	
	MIDAS	0.007	0.070	0.018	0.017	0.018	-	-	-	-	0.000	-	0.149	-	0.149	0.044	0.084	0.000	
	H-MIDAS	0.006	0.088	0.021	0.019	0.014	-	-	-	-	0.001	0.002	0.103	0.035	0.101	0.032	0.059	0.000	
CON	MEM	0.009	1.732	0.019	0.017	-	-	-	-	-	-	-	-	-	0.166	0.070	0.118	0.000	
	CMEM	0.008	0.709	0.017	0.017	0.030	0.032	0.052	-	-	-	-	-	-	0.170	0.076	0.115	0.000	
	HAR	0.009	0.341	0.019	0.018	0.066	-	0.039	0.063	0.084	-	-	-	-	0.149	0.059	0.086	0.000	
	MIDAS	0.008	0.485	0.018	0.017	0.020	-	-	-	-	0.000	-	39.039	-	0.129	0.054	0.081	0.000	
	H-MIDAS	0.008	0.097	0.021	0.019	0.039	-	-	-	-	0.001	0.002	0.073	0.974	0.194	0.090	0.136	0.000	
DTE	MEM	0.007	0.234	0.014	0.014	-	-	-	-	-	-	-	-	-	0.102	0.068	0.072	0.000	
	CMEM	0.007	0.337	0.014	0.015	0.011	0.020	0.026	-	-	-	-	-	-	0.100	0.068	0.079	0.000	
	HAR	0.007	0.303	0.014	0.015	0.038	-	0.031	0.052	0.065	-	-	-	-	0.079	0.048	0.059	0.000	
	MIDAS	0.007	0.858	0.016	0.016	0.016	-	-	-	-	0.000	-	21.410	-	0.088	0.064	0.071	0.000	
	H-MIDAS	0.006	0.577	0.017	0.017	0.016	-	-	-	-	0.001	0.002	11.768	61.903	0.059	0.041	0.043	0.000	
G1A	MEM	0.007	0.023	0.016	0.017	-	-	-	-	-	-	-	-	-	0.145	0.038	0.073	0.000	
	CMEM	0.007	0.038	0.018	0.018	0.003	0.017	0.019	-	-	-	-	-	-	0.272	0.064	0.131	0.000	
	HAR	0.007	0.036	0.018	0.018	0.024	-	0.035	0.060	0.066	-	-	-	-	0.117	0.028	0.057	0.000	
	MIDAS	0.007	0.049	0.021	0.019	0.014	-	-	-	-	0.000	-	0.105	-	0.134	0.031	0.064	0.000	
	H-MIDAS	0.007	0.056	0.022	0.020	0.013	-	-	-	-	0.001	0.002	5.704	5.693	0.078	0.019	0.037	0.000	
SZG	MEM	0.008	0.026	0.019	0.019	-	-	-	-	-	-	-	-	-	0.175	0.086	0.179	0.000	
	CMEM	0.007	0.046	0.021	0.020	0.015	0.028	0.042	-	-	-	-	-	-	0.082	0.039	0.080	0.000	
	HAR	0.007	0.051	0.021	0.021	0.040	-	0.040	0.061	0.073	-	-	-	-	0.314	0.146	0.315	0.000	
	MIDAS	0.007	0.069	0.023	0.021	0.019	-	-	-	-	0.001	-	5.474	-	0.024	0.014	0.034	0.000	
	H-MIDAS	0.007	0.074	0.025	0.022	0.021	-	-	-	-	0.001	0.002	0.143	5.656	0.027	0.018	0.010	0.000	
VOW	MEM	0.008	0.016	0.016	0.017	-	-	-	-	-	-	-	-	-	0.090	0.025	0.053	0.000	
	CMEM	0.007	0.029	0.017	0.018	0.005	0.024	0.028	-	-	-	-	-	-	0.094	0.025	0.056	0.000	
	HAR	0.007	0.028	0.017	0.018	0.019	-	0.028	0.049	0.050	-	-	-	-	0.093	0.024	0.055	0.000	
	MIDAS	0.007	0.037	0.019	0.018	0.012	-	-	-	-	0.000	-	0.030	-	0.104	0.027	0.061	0.000	
	H-MIDAS	0.007	0.041	0.021	0.019	0.009	-	-	-	-	0.001	0.002	3.546	3.526	0.066	0.017	0.038	0.000	

Table 12: In sample parameter estimates using the DZAF distribution: standard errors.

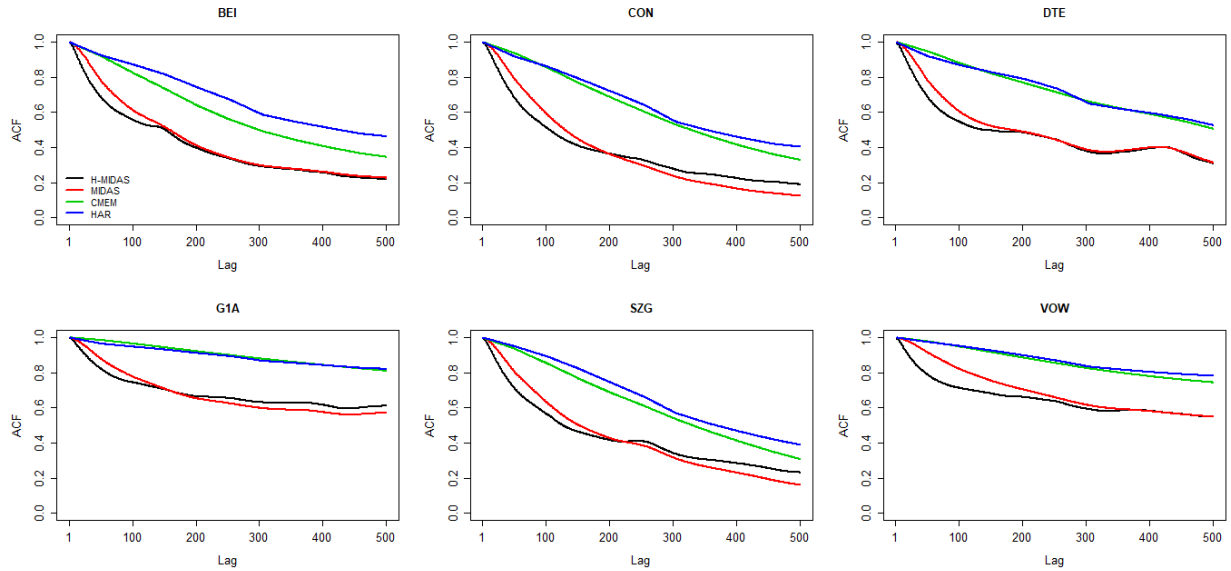
Ticker	Model	Intra-daily parameters					Trend parameters										ZAF parameters				
		α_1	α_0	β_1	β_2	m	$\alpha_{1,d}$	$\beta_{1,d}$	$\beta_{1,w}$	$\beta_{1,m}$	θ_d	θ_h	ω_d	ω_h	a	b	c	ϖ	δ_1	γ_1	
BEI	MEM	0.007	0.042	0.016	0.016	-	-	-	-	-	-	-	-	-	0.128	0.040	0.075	0.097	0.011	0.013	
	CNEM	0.007	0.058	0.017	0.017	0.016	0.022	0.034	-	-	-	-	-	-	0.096	0.031	0.058	0.106	0.011	0.014	
	HAR	0.007	0.059	0.017	0.017	0.037	-	0.033	0.045	0.054	-	-	-	-	0.110	0.035	0.065	0.110	0.011	0.015	
	MIDAS	0.006	0.078	0.019	0.017	0.017	-	-	-	-	0.000	-	0.142	-	0.102	0.032	0.061	0.100	0.011	0.014	
CON	H-MIDAS	0.006	0.092	0.021	0.018	0.014	-	-	-	-	0.001	0.002	7.142	51.112	0.128	0.039	0.075	0.018	0.011	0.003	
	MEM	0.008	0.496	0.015	0.016	-	-	-	-	-	-	-	-	-	0.191	0.082	0.131	0.099	0.016	0.012	
	CNEM	0.008	0.448	0.017	0.017	0.021	0.030	0.043	-	-	-	-	-	-	0.120	0.049	0.074	0.108	0.016	0.013	
	HAR	0.008	0.516	0.017	0.017	0.066	-	0.037	0.064	0.084	-	-	-	-	0.147	0.060	0.091	0.143	0.016	0.017	
DTE	MIDAS	0.008	0.469	0.018	0.018	0.030	-	-	-	-	0.001	-	0.220	-	0.119	0.049	0.074	0.162	0.017	0.020	
	H-MIDAS	0.008	0.668	0.020	0.019	0.040	-	-	-	-	0.001	0.002	5.879	0.647	0.075	0.030	0.055	0.172	0.022	0.019	
	MEM	0.007	0.350	0.014	0.014	-	-	-	-	-	-	-	-	-	0.086	0.060	0.065	0.854	0.189	0.033	
	CNEM	0.007	0.384	0.014	0.015	0.012	0.019	0.026	-	-	-	-	-	-	0.181	0.122	0.143	0.349	0.072	0.021	
GIA	HAR	0.007	0.471	0.014	0.015	0.037	-	0.030	0.051	0.061	-	-	-	-	0.030	0.027	0.033	0.261	0.269	0.023	
	MIDAS	0.007	0.344	0.016	0.016	0.020	-	-	-	-	0.001	-	0.196	-	0.030	0.026	0.030	0.138	0.136	0.016	
	H-MIDAS	0.007	0.244	0.015	0.014	0.019	-	-	-	-	0.001	0.002	27.509	1.695	0.096	0.066	0.073	0.058	0.011	0.010	
	MEM	0.008	0.029	0.017	0.018	-	-	-	-	-	-	-	-	-	0.137	0.033	0.063	0.055	0.014	0.009	
SZG	CNEM	0.007	0.042	0.019	0.018	0.006	0.035	0.040	-	-	-	-	-	-	0.124	0.030	0.061	0.074	0.015	0.011	
	HAR	0.007	0.040	0.018	0.018	0.024	-	0.035	0.060	0.067	-	-	-	-	0.136	0.032	0.065	0.066	0.015	0.010	
	MIDAS	0.007	0.053	0.020	0.019	0.014	-	-	-	-	0.000	-	0.073	-	0.127	0.030	0.062	0.066	0.018	0.010	
	H-MIDAS	0.007	0.060	0.022	0.019	0.013	-	-	-	-	0.001	0.002	0.061	0.169	0.125	0.029	0.059	0.054	0.014	0.008	
VOW	MEM	0.008	0.027	0.019	0.019	-	-	-	-	-	-	-	-	-	0.101	0.043	0.082	0.079	0.016	0.011	
	CNEM	0.007	0.053	0.021	0.020	0.016	0.028	0.042	-	-	-	-	-	-	0.103	0.048	0.101	0.088	0.018	0.013	
	HAR	0.008	0.050	0.021	0.021	0.040	-	0.040	0.062	0.072	-	-	-	-	0.110	0.051	0.106	0.081	0.017	0.012	
	MIDAS	0.007	0.077	0.023	0.021	0.019	-	-	-	-	0.001	-	0.105	-	0.095	0.044	0.089	0.085	0.018	0.012	
H-MIDAS	H-MIDAS	0.007	0.083	0.025	0.022	0.021	-	-	-	-	0.001	0.002	0.142	0.145	0.106	0.047	0.097	0.085	0.018	0.013	
	MEM	0.008	0.017	0.016	0.017	-	-	-	-	-	-	-	-	-	0.092	0.026	0.056	0.062	0.017	0.011	
	CNEM	0.007	0.031	0.017	0.017	0.005	0.024	0.027	-	-	-	-	-	-	0.097	0.025	0.057	0.073	0.019	0.013	
	HAR	0.007	0.034	0.017	0.018	0.019	-	0.029	0.049	0.051	-	-	-	-	0.101	0.026	0.060	0.066	0.019	0.012	
H-MIDAS	MIDAS	0.007	0.037	0.019	0.019	0.012	-	-	-	-	0.000	-	47.151	-	0.104	0.026	0.061	0.073	0.018	0.013	
	H-MIDAS	0.006	0.043	0.010	0.010	0.009	-	-	-	-	0.001	0.002	3.503	15.729	0.068	0.018	0.040	0.068	0.018	0.012	

Figure 11: Models with DZAF errors. Fitted long-run components vs seasonally adjusted volumes over a subsample of 5000 observations for BEI and VOW.



Key to figure: CMEM (green), HAR-CMEM (blue), MIDAS-CMEM (red) and H-MIDAS-CMEM (black). Seasonally adjusted intra-daily volumes are drawn in grey (for each series values are normalised with respect to the maximum observed trading volume).

Figure 12: Models with DZAF errors. Sample ACFs of the long-run components up to lag 500.



Key to figure: CMEM (green), HAR-CMEM (blue), MIDAS-CMEM (red) and H-MIDAS-CMEM (black). Models have been estimated by the ZAF distribution.

Table 13: Fitted H-MIDAS-CMEM models with DZAF errors: number of days taken by the Beta weights of the “daily” filter ($\phi_k(\omega_d)$) and number of 2 hour 50 minute - intervals taken by the Beta weights of the “hourly” filter ($\phi_{l,k}(\omega_h)$) to decay to 10^{-2} and 10^{-6} , respectively.

	Decay time to 10^{-2}		Decay time to 10^{-6}	
	daily	hourly	daily	hourly
BEI	7	5	23	14
CON	26	7	126	21
DTE	12	5	41	16
G1A	31	6	151	19
SZG	36	7	209	21
VOW	17	4	66	13

Figure 13: H-MIDAS-CMEM models with DZAF errors. Sample autocorrelation functions of the components $g_{t,i}$ (in red) and $\tau_{t,i}$ (in black) (up to 1 day).

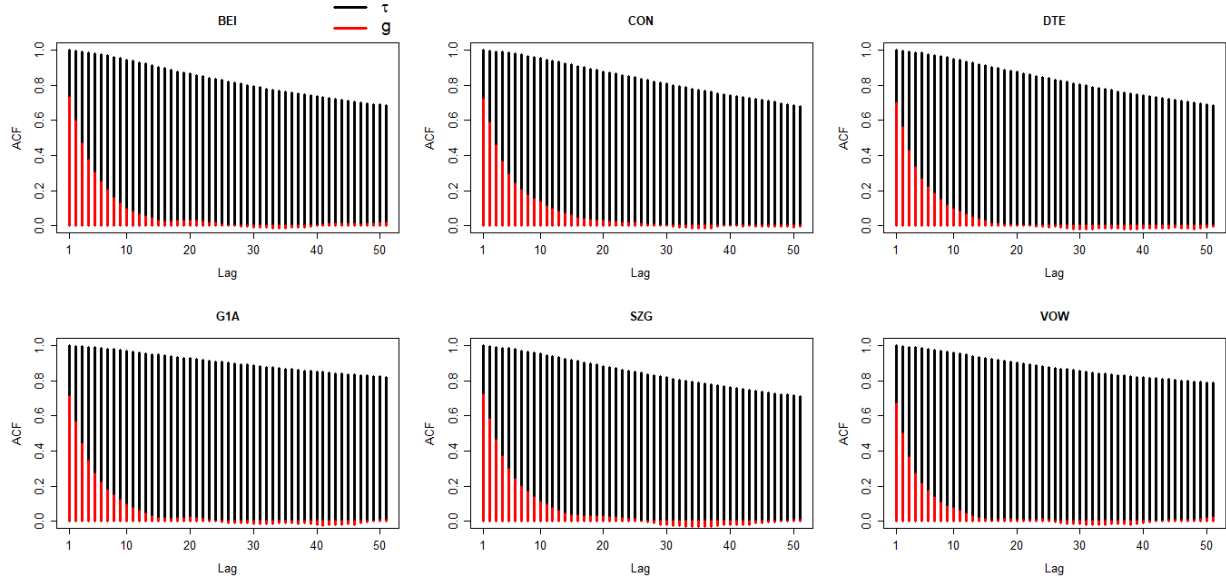


Figure 14: Decay profile of the Beta weighting function of the daily MIDAS filter for the H-MIDAS-CMEM with DZAF errors: beta weights (vertical axis) vs. daily lags (horizontal axis).

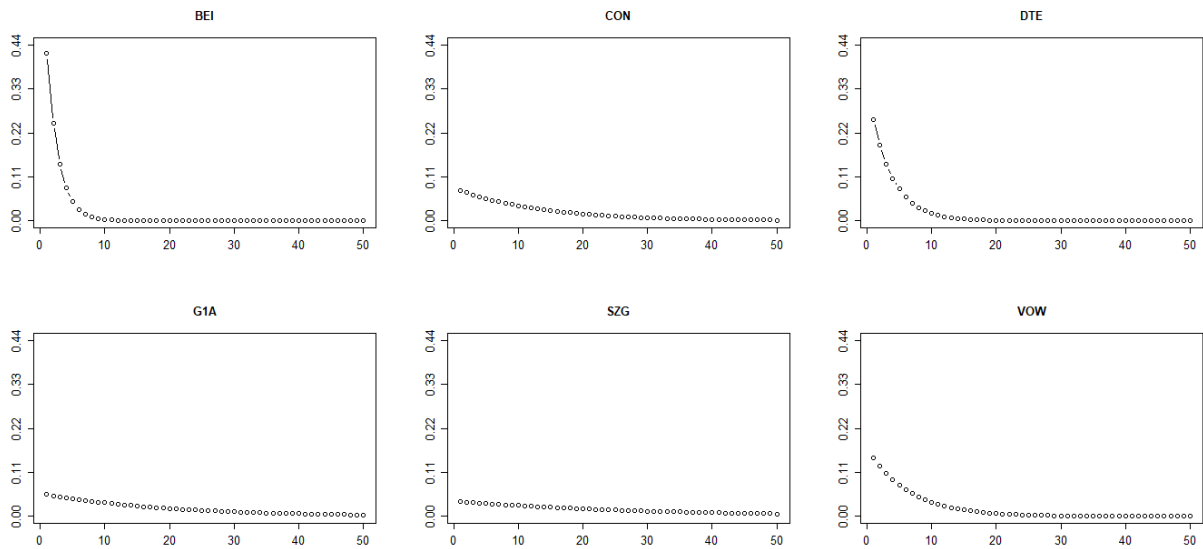


Figure 15: Decay profile of the Beta weighting function of the hourly MIDAS filter for the H-MIDAS-CMEM with DZAF errors: beta weights (vertical axis) vs. hourly lags (horizontal axis).

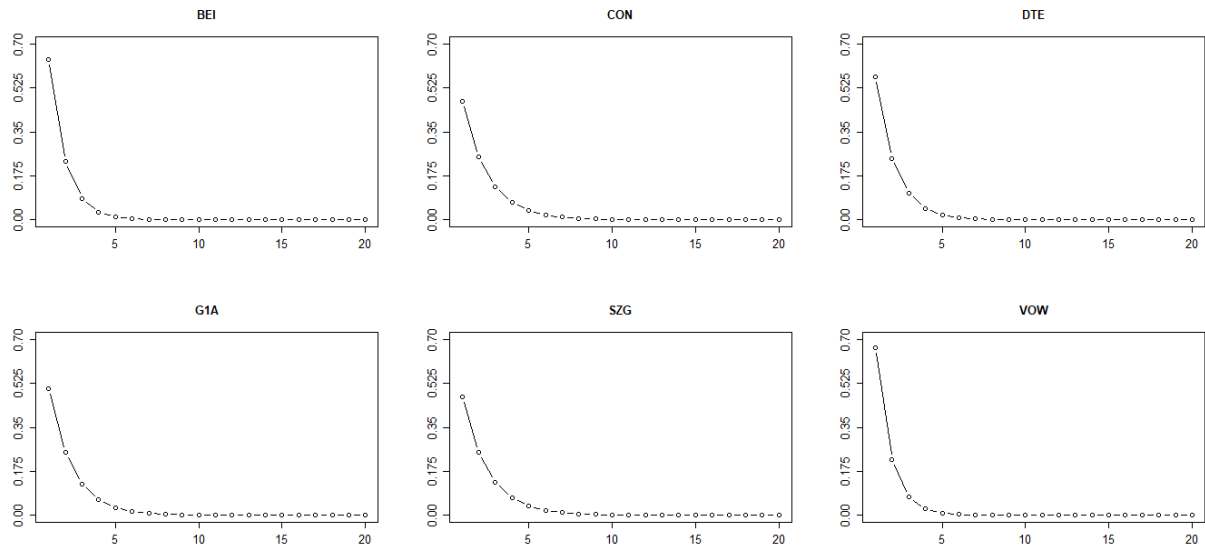


Table 14: Out-of-sample loss functions comparison for ZAF and DZAF: MCS p-values

	MCS p-values											
	BEI				CON				DTE			
	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>
MEM-ZAF	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.005	0.000	0.000	0.000
MEM-DZAF	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.004	0.000	0.000	0.000
CMEM-ZAF	0.000	0.000	0.000	0.000	0.007	0.000	0.001	0.005	0.007	0.271	0.972	1.000
CMEM-DZAF	0.000	0.000	0.000	0.000	0.007	0.000	0.000	0.001	0.008	0.000	0.000	0.001
HAR-ZAF	0.000	0.000	0.000	0.000	0.007	0.000	0.000	0.001	0.016	0.676	1.000	0.973
HAR-DZAF	0.000	0.000	0.000	0.000	0.008	0.000	0.001	0.004	0.005	0.000	0.000	0.000
MIDAS-ZAF	0.000	0.000	0.000	0.000	0.007	0.000	0.000	0.000	0.092	0.000	0.002	0.008
MIDAS-DZAF	0.000	0.000	0.000	0.000	0.012	0.018	0.047	0.069	0.021	0.000	0.022	0.062
H-MIDAS-ZAF	1.000	1.000	1.000	1.000	0.899	0.266	0.094	0.101	1.000	0.676	0.972	0.973
H-MIDAS-DZAF	0.023	0.062	0.095	0.099	1.000	1.000	1.000	1.000	0.441	1.000	0.972	0.973

	G1A				SZG				VOW			
	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>
	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>	<i>MSE</i>	<i>MAE</i>	<i>SL_{stc}</i>	<i>SL_{dyn}</i>
MEM-ZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MEM-DZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CMEM-ZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CMEM-DZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HAR-ZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HAR-DZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MIDAS-ZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MIDAS-DZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
H-MIDAS-ZAF	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.001	0.507	0.000	0.000
H-MIDAS-DZAF	0.385	0.513	0.482	0.433	0.019	0.027	0.000	0.000	1.000	1.000	1.000	1.000

The table shows the MCS p-values referring to the comparison in Table 8. In box model \in 75% MCS. Loss functions: Mean Squared Error (*MSE*), Mean Absolute Error (*MAE*) and *Slicing* loss with weights computed under the static (*SL_{stc}*) and dynamic (*SL_{dyn}*) VWAP replication strategy.

Table 15: Out-of-sample MSE comparison for different time horizons for ZAF and DZAF: MCS p-values

MCS p-values for MSE															
	BEI					CON					DTE				
	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30
MEM-ZAF	0.003	0.028	0.094	0.002	0.000	0.047	0.077	0.070	0.024	0.000	0.042	0.087	0.075	0.008	0.024
MEM-DZAF	0.005	0.038	0.094	0.002	0.000	0.046	0.089	0.072	0.024	0.000	0.042	0.165	0.216	0.025	0.024
CMEM-ZAF	0.027	0.068	0.076	0.006	0.000	0.059	0.089	0.072	0.024	0.000	0.379	0.165	0.098	0.008	0.031
CMEM-DZAF	0.022	0.067	0.076	0.006	0.000	0.053	0.089	0.104	0.047	0.000	0.045	0.165	0.098	0.010	0.076
HAR-ZAF	0.014	0.057	0.076	0.002	0.000	0.055	0.089	0.072	0.024	0.000	0.045	0.165	0.216	0.025	0.109
HAR-DZAF	0.012	0.055	0.076	0.002	0.000	0.053	0.089	0.072	0.044	0.000	0.042	0.165	0.115	0.013	0.076
MIDAS-ZAF	0.051	0.111	0.094	0.002	0.001	0.053	0.089	0.072	0.024	0.000	0.045	0.165	0.115	0.013	0.024
MIDAS-DZAF	0.060	0.111	0.094	0.002	0.000	0.104	0.164	0.135	0.047	0.000	0.131	0.165	0.115	0.040	0.078
H-MIDAS-ZAF	1.000	0.374	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
H-MIDAS-DZAF	0.878	1.000	0.296	0.100	0.820	0.648	0.387	0.330	0.483	0.100	0.379	0.351	0.216	0.088	0.741

	G1A					SZG					VOW				
	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30
MEM-ZAF	0.000	0.001	0.001	0.005	0.000	0.000	0.009	0.331	0.151	0.001	0.000	0.000	0.002	0.000	0.000
MEM-DZAF	0.000	0.001	0.001	0.004	0.000	0.000	0.024	0.989	0.351	0.001	0.000	0.000	0.001	0.000	0.000
CMEM-ZAF	0.001	0.031	0.003	0.005	0.000	0.001	0.277	0.128	0.086	0.001	0.000	0.001	0.027	0.000	0.000
CMEM-DZAF	0.001	0.044	0.003	0.005	0.000	0.001	0.107	0.192	0.086	0.001	0.000	0.000	0.032	0.000	0.000
HAR-ZAF	0.001	0.010	0.003	0.011	0.000	0.000	0.691	0.197	0.086	0.001	0.000	0.000	0.014	0.000	0.000
HAR-DZAF	0.001	0.015	0.003	0.006	0.000	0.000	0.148	0.128	0.086	0.001	0.000	0.000	0.014	0.000	0.000
MIDAS-ZAF	0.001	0.178	0.003	0.005	0.098	0.004	0.531	0.203	0.086	0.001	0.000	0.005	0.014	0.000	0.000
MIDAS-DZAF	0.001	0.495	0.003	0.006	0.028	0.002	0.691	0.157	0.053	0.001	0.000	0.009	0.010	0.000	0.000
H-MIDAS-ZAF	0.935	0.689	0.939	1.000	0.643	1.000	0.691	1.000	1.000	1.000	1.000	1.000	1.000	0.228	0.427
H-MIDAS-DZAF	1.000	1.000	1.000	0.609	1.000	0.799	1.000	0.331	0.151	0.269	0.235	0.406	0.601	1.000	1.000

The table shows the MCS p-values referring to the MSE comparison in Table 9. In box model \in 75% MCS. The results are based on the use of ZAF and DZAF distribution. Time horizons: 9:30 = 9:00 + 9:30; 10:00 = 9:30 + 10:00; 11:00 = 10:00 + 11:00; 13:00 = 11:00 + 13:00; 17:30 = 13:00 + 17:30.

Table 16: Out-of-sample MAE comparison for different time horizons for ZAF and DZAF: MCS p-values

Relative gain vs MEM-ZAF for MSE															
	BEI					CON					DTE				
	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30
MEM-ZAF	0.000	0.000	0.000	0.000	0.000	0.005	0.191	0.245	0.070	0.005	0.000	0.000	0.000	0.000	0.000
MEM-DZAF	0.000	0.000	0.000	0.000	0.000	0.009	0.807	1.000	0.853	0.723	0.000	0.000	0.000	0.000	0.000
CMEM-ZAF	0.000	0.000	0.000	0.000	0.000	0.024	0.807	0.961	0.853	0.078	1.000	1.000	1.000	0.440	0.087
CMEM-DZAF	0.000	0.000	0.000	0.000	0.000	0.007	0.532	0.739	0.508	0.000	0.000	0.000	0.000	0.000	0.000
HAR-ZAF	0.000	0.000	0.000	0.000	0.000	0.024	0.807	0.961	0.726	0.723	0.000	0.255	0.749	1.000	0.103
HAR-DZAF	0.000	0.000	0.000	0.000	0.000	0.008	0.573	0.746	0.646	0.029	0.000	0.000	0.000	0.000	0.000
MIDAS-ZAF	0.000	0.014	0.016	0.000	0.000	0.002	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MIDAS-DZAF	0.000	0.008	0.058	0.000	0.000	0.040	0.735	0.798	0.756	0.528	0.000	0.001	0.000	0.006	0.000
H-MIDAS-ZAF	1.000	1.000	0.947	1.000	0.734	0.608	0.807	0.961	0.853	1.000	0.000	0.013	0.020	0.101	0.024
H-MIDAS-DZAF	0.201	0.784	1.000	0.095	1.000	1.000	1.000	0.961	1.000	0.958	0.000	0.006	0.004	0.266	1.000

	G1A					SZG					VOW				
	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30	9:30	10:00	11:00	13:00	17:30
MEM-ZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MEM-DZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CMEM-ZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.039	0.012	0.000	0.000	0.000	0.000	0.000	0.000
CMEM-DZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000
HAR-ZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.018	0.003	0.000	0.000	0.000	0.000	0.000	0.000
HAR-DZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000
MIDAS-ZAF	0.000	0.000	0.000	0.000	0.000	0.006	0.025	0.142	0.012	0.000	0.000	0.000	0.000	0.000	0.000
MIDAS-DZAF	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.039	0.001	0.000	0.000	0.000	0.000	0.000	0.000
H-MIDAS-ZAF	1.000	1.000	1.000	1.000	0.829	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
H-MIDAS-DZAF	0.425	0.306	0.211	0.275	1.000	0.065	0.025	0.059	0.012	0.145	0.119	0.255	0.276	0.792	0.821

The table shows the MCS p-values referring to the MAE comparison in Table 10. In box model \in 75% MCS. The results are based on the use of ZAF and DZAF distribution. Time horizons: 9:30 = 9:00 + 9:30; 10:00 = 9:30 + 10:00; 11:00 = 10:00 + 11:00; 13:00 = 11:00 + 13:00; 17:30 = 13:00 + 17:30.

References

- Amado, C., A. Silvennoinen, and T. Terasvirta (2019). *Models with Multiplicative Decomposition of Conditional Variances and Correlations*, Volume 2. United Kingdom: Routledge.
- Amado, C. and T. Teräsvirta (2013). Modelling volatility by variance decomposition. *Journal of Econometrics* 175(2), 142–153.
- Andersen, T. G., T. Bollerslev, and J. Cai (2000). Intraday and interday volatility in the Japanese stock market. *Journal of International Financial Markets, Institutions and Money* 10(2), 107–130.
- Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica* 71(2), 579–625.
- Bauwens, L., M. Braione, and G. Storti (2016). Forecasting comparison of long term component dynamic models for realized covariance matrices. *Annals of Economics and Statistics* (123/124), 103–134.
- Bauwens, L., M. Braione, and G. Storti (2017). A dynamic component model for forecasting high-dimensional realized covariance matrices. *Econometrics and Statistics* 1(C), 40–61.
- Berkowitz, S. A., D. E. Logue, and E. A. Noser (1988). The total cost of transactions on the NYSE. *The Journal of Finance* 43(1), 97–112.
- Bougerol, P. and N. Picard (1992a). Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics* 52(1), 115–127.
- Bougerol, P. and N. Picard (1992b). Strict stationarity of generalized autoregressive processes. *The Annals of Probability* 20(4), 1714–1730.
- Brownlees, C. T., F. Cipollini, and G. M. Gallo (2011). Intra-daily volume modeling and prediction for algorithmic trading. *Journal of Financial Econometrics* 9(3), 489–518.
- Brownlees, C. T. and G. M. Gallo (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis* 51(4), 2232–2245.
- Brownlees, C. T. and G. M. Gallo (2010). Comparison of volatility measures: a risk management perspective. *Journal of Financial Econometrics* 8(1), 29–56.
- Brunetti, C. and P. M. Lildholdt (2007). Time series modeling of daily log-price ranges for CHF/USD and USD/GBP. *The Journal of Derivatives* 15(2), 39–59.
- Chou, R. Y. (2005). Forecasting financial volatilities with extreme values: the conditional autoregressive range (CARR) model. *Journal of Money, Credit and Banking*, 561–582.
- Cipollini, F., R. F. Engle, and G. M. Gallo (2006). Vector multiplicative error models: representation and inference. Technical report, National Bureau of Economic Research.
- Cipollini, F., R. F. Engle, and G. M. Gallo (2013). Semiparametric vector mem. *Journal of Applied Econometrics* 28(7), 1067–1086.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 174–196.
- Engle, R. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics* 17(5), 425–446.
- Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics* 95(3), 776–797.
- Engle, R. F. and J. G. Rangel (2008). The spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *Review of Financial Studies* 21(3), 1187–1222.
- Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 1127–1162.
- Engle, R. F. and M. E. Sokalska (2012). Forecasting intraday volatility in the US equity market. multiplicative component GARCH. *Journal of Financial Econometrics* 10(1), 54–83.
- Franco, C. and J. Zakoian (2011). *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley.
- Gallant, A. R. (1981). On the bias in flexible functional forms and an essentially unbiased form: the Fourier flexible form. *Journal of Econometrics* 15(2), 211–245.
- Gallo, G. M. and E. Otranto (2015). Forecasting realized volatility with changing average levels. *International Journal of Forecasting* 31(3), 620–634.
- Ghysels, E., P. Santa-Clara, and R. Valkanov (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics* 131(1), 59–95.
- Ghysels, E., A. Sinko, and R. Valkanov (2007). Midas regressions: Further results and new directions. *Econometric Reviews* 26(1), 53–90.
- Glasserman, P. and D. D. Yao (1995). Stochastic vector difference equations with stationary coefficients. *Journal of Applied Probability* 32(4), 851–866.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50(4), 1029–1054.
- Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica* 79(2), 453–497.
- Hautsch, N. (2003). Assessing the risk of liquidity suppliers on the basis of excess demand intensities. *Journal of Financial Econometrics* 1(2), 189–215.
- Hautsch, N., P. Malec, and M. Schienle (2014). Capturing the zero: A new class of zero-augmented distributions and multiplicative error processes. *Journal of Financial Econometrics* 12(1), 89–121.
- Lanne, M. (2006). A mixture multiplicative error model for realized volatility. *Journal of Financial Econometrics* 4(4), 594–616.
- Lee, S.-W. and B. E. Hansen (1994). Asymptotic theory for the GARCH (1, 1) quasi-maximum likelihood estimator. *Econometric theory* 10(01), 29–52.
- Love, E. R. (1980). 64.4 some logarithm inequalities. *The Mathematical Gazette* 64(427), 55–57.
- Madhavan, A. N. (2002). VWAP strategies. *Trading* 2002(1), 32–39.
- Manganelli, S. (2005). Duration, volume and volatility impact of trades. *Journal of Financial Markets* 8(4), 377–399.
- Müller, U. A., M. M. Dacorogna, R. D. Davé, O. V. Pictet, R. B. Olsen, and J. R. Ward (1993). Fractals and intrinsic time: A challenge to econometricians. *Unpublished manuscript, Olsen & Associates, Zürich*.
- Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. Volume 4 of *Handbook of Econometrics*, pp. 2111–2245. Elsevier.

- Patton, A., D. N. Politis, and H. White (2009). Correction to "automatic block-length selection for the dependent bootstrap" by d. politis and h. white. *Econometric Reviews* 28(4), 372–375.
- Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica* 39(3), 577–591.
- Russell, J. R. and R. F. Engle (2005). A discrete-state continuous-time model of financial transactions prices and times: The autoregressive conditional multinomial–autoregressive conditional duration model. *Journal of Business & Economic Statistics* 23(2), 166–180.
- Wang, F. and E. Ghysels (2015). Econometric analysis of volatility component models. *Econometric Theory* 31(2), 362–393.